17th International Symposium on Hearing

15-19 June 2015, Groningen, Netherlands

Editors:

Pim van Dijk
Deniz Başkent
Etienne Gaudrain
Emile de Kleine
Anita Wagner
Cris Lanting

University Medical Center Groningen, Department of Otorhinolaryngology, Groningen, Netherlands

University of Groningen, Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, Groningen, Netherlands

This edition of the ISH was supported by:

- The Royal Netherlands Academy of Arts and Sciences (KNAW)
- The Oticon Fonden
- The Heinsius-Houbolt Fonds
- Phonak
- Med-El
- Cochlear
- The City of Groningen
- The Province of Groningen
- The Research School of Behavioral and Cognitive Neurosciences
- The University of Groningen
- The University Medical Center Groningen

Version 15/06/2015

Table of Contents

Chapter 1. Effects of Age and Hearing Loss on the Processing of Auditory Temporal Fine Structure — *B. C. J. Moore* 7

Chapter 2. Aging Effects on Behavioural Estimates of Suppression with Short Suppressors — *E. Hegland, E. Strickland* 15

Chapter 3. Contributions of Coding Efficiency of Temporal-Structure and Level Information to the Lateralization Performance in Young and Early-Elderly Listeners — *A. Ochi, T. Yamasoba, S. Furukawa* 23

Chapter 4. Investigating the Role of Working Memory in Speech-in-Noise Identification for Listeners with Normal Hearing — *C. Füllgrabe, S. Rosen* 31

Chapter 5. The Contribution of Auditory and Cognitive Factors to Intelligibility of Words and Sentences in Noise — *A. Heinrich, S. Knight* 39

Chapter 6. Do Hearing Aids Improve Affect Perception? — *J. Schmidt, D. Herzog, O. Scharenborg, E. Janse* 47

Chapter 7. Suitability of the Binaural Interaction Component for Interaural Electrode Pairing of Bilateral Cochlear Implants — *H. Hu, M. Dietz, B. Kollmeier* 55

Chapter 8. Binaural Loudness Constancy — *J. Culling, H. Dare* 63

Chapter 9. Intelligibility for Binaural Speech with Discarded Low-SNR Speech Components — *E. Schoenmaker, S. van de Par* 69

Chapter 10. On the Contribution of Target Audibility to Performance in Spatialized Speech Mixtures — *V. Best, C. R. Mason, J. Swaminathan, G. Kidd, Jr., K. M. Jakien, S. D. Kampel, F. J. Gallun, J. M. Buchholz, H. Glyde* 77

Chapter 11. Optimization of a Spectral Contrast Enhancement Algorithm for Cochlear Implants based on a Vowel Identification Model — *W. Nogueira Vazquez, T. Rode, A. Büchner* . 85

Chapter 12. Roles of the Contralateral Efferent Reflex in Hearing Demonstrated with Cochlear Implants — *E. A. Lopez-Poveda, A. Eustaquio-Martin, J. S. Stohl, R. D. Wolford, R. Schatzer, B. S. Wilson* 93

Chapter 13. Deactivating Cochlear Implant Electrodes Based on Pitch Information for Users of the ACE Strategy — *D. Vickers, A. Degun, A. Canas, T. Stainsby, F. Vanpoucke* 101

Chapter 14. Speech Masking and Masking Release in Normal, Impaired and Electric Hearing: Interactions Between Frequency Selectivity and Inherent Temporal Fluctuations in Noise — *A. J. Oxenham, H. A. Kreft* 109

Chapter 15. Effects of Pulse Shape and Polarity on Sensitivity to Cochlear Implant Stimulation: a Chronic Study in Guinea Pigs — *O. Macherey, Y. Cazals* 117

Chapter 16. Assessing the Firing Properties of the Electrically Stimulated Auditory Nerve using a Convolution Model — *S. B. Strahl, D. Ramekers, M. M. B. Nagelkerke, K. E. Schwarz, P. Spitzer, S. F. L. Klis, W. Grolman, H. Versnel* 127

Chapter 17. Modeling the Individual Variability of Loudness Perception with a Multi-Category Psychometric Function — *A. C. Trevino, W. Jesteadt, S. T. Neely* 135

Chapter 18. Auditory fMRI of Sound Intensity and Loudness for Unilateral Stimulation — *O. Behler, S. Uppenkamp* 143

Chapter 19. Tinnitus-related Abnormalities in Brain Networks During Stimulated and Resting States — *C. Lanting, A. Woźniak, P. van Dijk, D. Langers* 153

Chapter 20. The Role of Conduction Delay in Creating Sensitivity to Interaural Time Differences — *C. E. Carr, G. Ashida, H. Wagner, T. McColgan, R. Kempter* 163

Chapter 21. Objective Measures of Neural Processing of Interaural Time Differences — *D. McAlpine, N. Haywood, J. A. Undurraga, T. Marquardt* 171

Chapter 22. Minimum Audible Angles measured with Simulated Normally-Sized and Oversized Pinnas for Normal-Hearing and Hearing-Impaired Test Subjects — *F. M. Rønne, S. Laugesen, N. S. Jensen, J. H. Pedersen* 179

Chapter 23. Moving Objects in the Barn Owl's Auditory World — *U. Langemann, B. Krumm, K. Liebner, R. Beutelmann, G. M. Klump* 187

Chapter 24. Change Detection in Auditory Textures — *Y. Boubenec, J. Lawlor, S. Shamma, B. Englitz* 195

Chapter 25. The Relative Contributions of Temporal Envelope and Fine Structure in Mandarin Lexical Tone Perception in Auditory Neuropathy Spectrum Disorder — *S. Wang, R. Dong, Y. Wang, D. Liu, L. Zhang, L. Xu* 203

Chapter 26. Interaction of Object Binding Cues in Binaural Masking Pattern Experiments — *J. Verhey, B. Lübken, S. van de Par* 209

Chapter 27. Speech Intelligibility for Target and Masker with a Different Spectrum — *T. Leclère, D. Thery, M. Lavandier, J. F. Culling* 217

Chapter 28. Dynamics of Cochlear Nonlinearity — *N. P. Cooper, M. van der Heijden* 225

Chapter 29. Responses of the Human Inner Ear to Low-Frequency Sound — *M. Drexl, E. Krause, R. Gürkov, L. Wiegrebe* 231

Chapter 30. Suppression Measured from Chinchilla Auditory-Nerve-Fiber Responses Following Noise-Induced Hearing Loss: Adaptive-Tracking and Systems-Identification Approaches — *M. Sayles, M. K. Walls, M. G. Heinz* 239

Chapter 31. Does Signal Degradation Affect Top-Down Processing of Speech? — *A. Wagner, C. Pals, C. de Blecourt, A. Sarampalis, D. Başkent* 249

Chapter 32. The Effect of Peripheral Compression on Syllable Perception Measured with a Hearing Impairment Simulator — *T. Matsui, T. Irino, M. Nagae, H. Kawahara, R. D. Patterson* 257

Chapter 33. Towards Objective Measures of Functional Hearing Abilities — *H. Innes-Brown, R. Tsongas, J. Marozeau, C. McKay* 265

Chapter 34. Connectivity in Language Areas of the Brain in Cochlear Implant Users as Revealed by fNIRS — *C. McKay, A. Shah, A. K. Seghouane, X. Zhou, L. Tate, W. Cross, R. Litovsky* 275

Chapter 35. Isolating Neural Indices of Continuous Speech Processing at the Phonetic Level — *G. Di Liberto, E. Lalor* 283

Chapter 36. Brainstem Coding of Pitch – the Butte Hypothesis — *P. X. Joris* 293

Chapter 37. Can Temporal Fine Structure and Temporal Envelope be Considered Independently for Pitch Perception? — *N. Grimault* 295

Chapter 38. Locating Melody Processing Activity in Auditory Cortex with MEG — *R. D. Patterson, M. Andermann, S. Uppenkamp, A. Rupp* 303

Chapter 39. Studying Effects of Transcranial Alternating Current Stimulation on Hearing and Auditory Scene Analysis — *L. Riecke* 309

Chapter 40. Functional Organization of the Ventral Auditory Pathway — *Y. E. Cohen, S. Bennur, K. L. Christison-Lagay, A. M. Gifford, J. Tsunada* 319

Chapter 41. Neural Segregation of Concurrent Speech: Effects of Background Noise and Reverberation on Auditory Scene Analysis in the Ventral Cochlear Nucleus — *M. Sayles, A. Stasiak, I. M. Winter* 327

Chapter 42. Audio Visual Integration with Competing Sources in the Framework of Audio Visual Speech Scene Analysis — *G. Attigodu Chandrashekara, F. Berthommier, J. L. Schwartz* 335

Chapter 43. Relative Pitch Perception and the Detection of Deviant Tone Patterns — *S. L. Denham, M. Coath, G. Háden, F. Murray, I. Winkler* 343

Chapter 44. Neural Correlates of Auditory-Tactile Integration in Musical Meter Perception — *J. Huang, T. Wu, S. Hsiao, X. Wang* 351

Chapter 45. Do Zwicker Tones Evoke a Musical Pitch? — *H. E. Gockel, R. P. Carlyon* 357

Chapter 46. Speech Coding in the Midbrain: Effects of Sensorineural Hearing Loss — *L. H. Carney, D. O. Kim, S. Kuwada* 363

Chapter 47. Sources of Variability in Consonant Perception and Implications for Speech Perception Modeling — *J. Zaar, T. Dau* 371

Chapter 48. On Detectable and Meaningful Speech-Intelligibility Benefits — *W. Whitmer, D. McShefferty, M. Akeroyd* 379

Chapter 49. Individual Differences in Behavioral Decision Weights Related to Irregularities in Cochlear Mechanics — *J. Lee, I. Heo, A. C. Chang, K. Bond, C. Stoelinga, R. Lutfi, G. R. Long* 387

Chapter 50. On the Interplay between Cochlear Gain Loss and Temporal Envelope Coding Deficits — *S. Verhulst, P. Piktel, A. Jagadeesh, M. Mauermann* 395

Chapter 51. Frequency Tuning of the Human Efferent Effect on Cochlear Gain — *V. Drga, C. J. Plack, I. Yasin* 403

Chapter 1

Effects of Age and Hearing Loss on the Processing of Auditory Temporal Fine Structure

Brian C.J. Moore

Department of Experimental Psychology, University of Cambridge, Cambridge, UK

Contact author: BCJM, bcjm@cam.ac.uk

Abstract

Within the cochlea, broadband sounds are filtered into a series of narrowband signals, each of which has a relatively slowly varying envelope (ENV) imposed on a rapidly oscillating carrier (the temporal fine structure, TFS). Information about ENV and TFS is conveyed in the timing and short-term rate of nerve spikes in the auditory nerve. This paper summarises evidence from studies that allow some separation of the effects of hearing loss and age on the processing of TFS and ENV information. The monaural processing of TFS information seems to be adversely affected by both hearing loss and increasing age. The monaural processing of ENV information is hardly affected by hearing loss or by increasing age. The binaural processing of TFS information is also adversely affected by both hearing loss and increasing age, but here the latter seems more important. The binaural processing of ENV information deteriorates somewhat with increasing age but is not markedly affected by hearing loss. The reduced TFS processing abilities found for older/hearing-impaired subjects may partially account for the difficulties that such subjects experience in complex listening situations.

Keywords: frequency discrimination, envelope, binaural processing, interaural phase discrimination, envelope processing, intelligibility, processing efficiency

1 Introduction

Within the healthy cochlea, broadband sounds are decomposed into narrowband signals, each of which can be considered as a relatively slowly varying envelope (ENV) imposed on a rapidly oscillating carrier (the temporal fine structure, TFS). Information about ENV and TFS is conveyed in the timing and short-term rate of nerve spikes in the auditory nerve (Joris & Yin, 1992). Following Moore (2014), a distinction is made between the physical ENV and TFS of the input signal (ENV_p and TFS_p), the ENV and TFS at a given place on the basilar membrane (ENV_{BM} and TFS_{BM}), and the neural representation of ENV_{BM} and TFS_{BM} (ENV_n and TFS_n). This paper reviews studies that separate the effects of hearing loss and age on the auditory processing of ENV and TFS.

2 Effects of Age

2.1 Monaural Processing of TFS

It is widely believed that the difference limen for frequency (DLF) of pure tones depends on the use of TFS information for frequencies up to 4-5 kHz (Moore, 1973), and perhaps even up to 8 kHz (Moore & Ernst, 2012). If so, the DLF at low and medium frequencies provides a measure of TFS processing. Abel et al. (1990) and He et al. (1998) compared DLFs for young and older subjects with normal hearing (audiometric thresholds ≤20 dB HL from 0.25 to 4 kHz). In both studies, DLFs were larger by a factor of 2-4 for the older than for the younger subjects, suggesting that there is an effect of age. However, performance on other tasks (duration discrimination for Abel et al. and intensity discrimination for He et al.) was also poorer for the older group, suggesting that there may be a general effect of age that leads to reduced "processing efficiency". The effect of age cannot be attributed to reduced frequency selectivity, since auditory filters do not broaden with increasing age when absolute thresholds remain normal (Lutman et al., 1991; Peters & Moore, 1992).

Monaural TFS processing has also been assessed using the TFS1 test (Hopkins & Moore, 2007; Moore & Sek, 2009). This requires discrimination of a harmonic complex tone (H) from an inharmonic tone (I), created by shifting all components in the H tone up by a fixed amount in hertz, Δf. The value of Δf is adaptively varied to determine the threshold. Both tones are passed through a fixed bandpass filter centred on the higher (unresolved) components, and a background noise is used to mask combination tones and components falling on the skirts of the filter. Evidence suggesting that performance on this test reflects sensitivity to TFS rather than the use of excitation-pattern cues is: (1) Performance on the TFS test does not worsen with increasing level, except when the level is very high (Moore & Sek, 2009; 2011; Marmel et al., 2015); (2) Randomizing the level of the individual components would be expected to impair the ability to use excitation-pattern cues, but this has very little effect on performance of the TFS1 test (Jackson & Moore, 2014); (3) Differences in excitation level between the H and I tones at the threshold value of Δf, estimated using the H and I tones as forward maskers, are too small to be usable (Marmel et al., 2015).

Moore et al. (2012) used the TFS1 test with centre frequencies of 850 and 2000 Hz to test 35 subjects with audiometric thresholds ≤20 dB HL from 0.25 to 6 kHz and ages from 22 to 61 years. There were significant correlations between age and the thresholds in the TFS1 test (r = 0.69 and 0.57 for the centre frequencies of 850 and 2000 Hz, respectively). However, TFS1 scores at 850 Hz were correlated with absolute thresholds at 850 Hz (r = 0.67), even though audiometric thresholds were within the normal range, making it hard to rule out an effect of hearing loss.

Füllgrabe et al. (2015) eliminated confounding effects of hearing loss by testing young (18-27 years) and older (60-79 years) subjects with *matched* audiograms. Both groups had audiometric thresholds ≤20 dB HL from 0.125 to 6 kHz. The TFS1 test was conducted using centre frequencies of 1 and 2 kHz. The older subjects performed significantly more poorly than the young subjects for both centre frequencies. Performance on the TFS1 task was not correlated with audiometric thresholds at the test frequencies. These results confirm that increased age is associated with a poorer ability to use monaural TFS cues, in the absence of any abnormality in the audiogram. It is hard to know whether this reflects a specific deficit in TFS processing or a more general reduction in processing efficiency.

2.2 Monaural Processing of ENV

Füllgrabe et al. (2015) also assessed sensitivity to ENV cues by measuring thresholds for detecting sinusoidal amplitude modulation imposed on a 4-kHz sinusoidal carrier. Modulation rates of 5, 30, 90, and 180 Hz were used to characterize the temporal-modulation-transfer function (TMTF). On average, thresholds (expressed as $20\log_{10}m$, where m is the modulation

index) were 2-2.5 dB higher (worse) for the older than for the younger subjects. However, the shapes of the TMTFs were similar for the two groups. This suggests that increasing age is associated with reduced efficiency in processing ENV information, but not with reduced temporal resolution for ENV cues. Schoof and Rosen (2014) found no significant difference in either processing efficiency or the shape of TMTFs measured with noise-band carriers between young (19-29 years) and older (60-72 years) subjects with near-normal audiograms. It is possible that, with noiseband carriers, amplitude-modulation detection is limited by the inherent fluctuations in the carrier (Dau et al., 1997), making it hard to measure the effects of age.

2.3 Binaural Processing of TFS

The binaural processing of TFS can be assessed by measuring the smallest detectable interaural phase difference (IPD) of a sinusoidal carrier relative to an IPD of 0°, keeping the envelope synchronous at the two ears. A task of this type is the TFS-LF test (Hopkins & Moore, 2010b; Sek & Moore, 2012). Moore et al. (2012) tested 35 subjects with audiometric thresholds ≤20 dB HL from 0.25 to 6 kHz and ages from 22 to 61 years. The TFS-LF test was used at center frequencies of 500 and 850 Hz. There were significant correlations between age and IPD thresholds at both 500 Hz (r = 0.37) and 850 Hz (r = 0.65). Scores on the TFS-LF task were not significantly correlated with absolute thresholds at the test frequency. These results confirm those of Ross et al. (2007) and Grose and Mamo (2010), and indicate that the decline in sensitivity to binaural TFS with increasing age is already apparent by middle age.

Füllgrabe et al. (2015) used the TFS-LF task with their young and older groups with matched (normal) audiograms, as described above. The older group performed significantly more poorly than the young group at both centre frequencies used (500 and 750 Hz). The scores on the TFS-LF test were not significantly correlated with audiometric thresholds at the test frequencies. Overall, the results indicate that the ability to detect changes in IPD worsens with increasing age even when audiometric thresholds remain within the normal range.

The binaural masking level difference (BMLD) provides another measure of binaural sensitivity to TFS. The BMLD is the difference in the detection threshold between a condition where the masker and signal have the same interaural phase and level relationship, and a condition where the relationship is different. Although BMLDs can occur as a result of differences in ENV_n for the two ears, it is generally believed that the largest BMLDs result from the use of TFS_n. Because the BMLD represents a *difference* between two thresholds, it has the advantage that differences in processing efficiency across age groups at least partly cancel. Several researchers have compared BMLDs for young and older subjects with (near-) normal hearing (Pichora-Fuller & Schneider, 1991; 1992; Grose et al., 1994; Strouse et al., 1998). All showed that BMLDs (usually for a 500-Hz signal frequency) were smaller for older than for young subjects, typically by 2-4 dB.

In summary, binaural TFS processing deteriorates with increasing age, even when audiometric thresholds are within the normal range.

2.4 Binaural Processing of ENV

King et al. (2014) measured the ability to discriminate IPD using amplitude-modulated sinusoids. The IPDs were imposed either on the carrier, TFS_p, or the modulator, ENV_p. The carrier frequency, f_c, was 250 or 500 Hz and the modulation frequency was 20 Hz. They tested 46 subjects with a wide range of ages (18 to 83 years) and degrees of hearing loss. The absolute thresholds at the carrier frequencies of 250 and 500 Hz were not significantly correlated with age. Thresholds for detecting IPDs in ENV_p were positively correlated with age for both carrier

frequencies (r = 0.62 for f_c = 250 Hz; r = 0.58, for f_c = 500 Hz). The correlations remained positive and significant when the effect of absolute threshold was partialled out. These results suggest that increasing age adversely affects the ability to discriminate IPDs in ENV_p, independently of the effects of hearing loss.

3 Effects of Cochlear Hearing Loss

3.1 Monaural Processing of TFS

Many studies have shown that cochlear hearing loss is associated with larger than normal DLFs. For a review, see Moore (2014). However, in most studies the hearing-impaired subjects were older than the normal-hearing subjects, so some (but probably not all) of the effects of hearing loss may have been due to age. DLFs for hearing-impaired subjects are not correlated with measures of frequency selectivity (Tyler et al., 1983; Moore & Peters, 1992), confirming that DLFs depend on TFS information and not excitation-pattern information.

Several studies have shown that cochlear hearing loss is associated with a greatly reduced ability to perform the TFS1 test or similar tests (Hopkins & Moore, 2007; 2010a; 2011). Many hearing-impaired subjects cannot perform the task at all. Although the effects of hearing loss and age were confounded in some of these studies, the performance of older hearing-impaired subjects seems to be much worse than that of older normal-hearing subjects, suggesting that hearing loss *per se* adversely affects monaural TFS processing.

3.2 Monaural Processing of ENV

Provided that the carrier is fully audible, the ability to detect amplitude modulation is generally not adversely affected by cochlear hearing loss and may sometimes be better than normal, depending on whether the comparison is made at equal sound pressure level or equal sensation level (Bacon & Viemeister, 1985; Bacon & Gleitman, 1992; Moore et al., 1992; Moore & Glasberg, 2001). When the modulation is clearly audible, a sound with a fixed modulation depth appears to fluctuate more for an impaired ear than for a normal ear, probably because of the effects of loudness recruitment (Moore et al., 1996).

3.3 Binaural Processing of TFS

Many studies of the effects of hearing loss on discrimination of interaural time differences (ITDs) are confounded by differences in age between the normal-hearing and hearing-impaired subjects. One exception is the study of Hawkins and Wightman (1980). They measured just noticeable differences in ITD using low-frequency (450-550 Hz) and high-frequency (3750-4250 Hz) narrow-band noise stimuli. The ITD was present in both TFS_p and ENV_p, except that the onsets and offsets of the stimuli were synchronous across the two ears. Performance was probably dominated by the use of TFS cues for the low frequency and ENV cues for the high frequency. They tested three normal-hearing subjects (mean age 25 years) and eight subjects with hearing loss (mean age 27 years). The two groups were tested both at the same sound pressure level (85 dB SPL) and at the same sensation level (30 dB SL). ITD thresholds were significantly higher for the hearing-impaired than for the normal-hearing subjects for both signals at both levels. These results suggest that cochlear hearing loss can adversely affect the discrimination of ITD carried both in TFS_p and in ENV_p.

The study of King et al. (2014), described above, is relevant here. Recall that they tested 46 subjects with a wide range of ages (18 to 83 years) and degrees of hearing loss. Thresholds for detecting IPDs in TFS_p were positively correlated with absolute thresholds for both carrier frequencies (r = 0.45 for f_c = 250 Hz; r = 0.40, for f_c = 500 Hz). The correlations remained positive and significant when the effect of age was partialled out. Thus, hearing loss adversely affects the binaural processing of TFS.

Many studies have shown that BMLDs are reduced for people with cochlear hearing loss, but again the results are generally confounded with the effects of age. More research is needed to establish the extent to which BMLDs are affected by hearing loss *per se*.

3.4 Binaural Processing of ENV

Although the results of Hawkins and Wightman (1980), described in section 3.3, suggested that hearing loss adversely affected the discrimination of ITD carried in ENV_p, King et al. (2014) found that thresholds for detecting IPDs in ENV_p were not significantly correlated with absolute thresholds. Also, Léger et al. (2015) reported that thresholds for detecting IPDs in ENV_p were similar for subjects with normal and impaired hearing (both groups had a wide range of ages). Overall, the results suggest that hearing loss does not markedly affect the ability to process binaural ENV cues.

4 Summary and Implications

The monaural and binaural processing of TFS is adversely affected by both increasing age and cochlear hearing loss. The efficiency of processing monaural ENV information may be adversely affected by increasing age, but the temporal resolution of ENV cues appears to be unaffected. The processing of binaural ENV information is adversely affected by increasing age. Cochlear hearing loss does not markedly affect the processing of monaural or binaural ENV information.

The effects of age and hearing loss on TFS processing may partially explain the difficulties experienced by older hearing-impaired people in understanding speech in background sounds (Hopkins & Moore, 2010a; 2011; Neher et al., 2012; Füllgrabe et al., 2015).

Acknowledgments

This work was supported by grant G0701870 from the MRC (UK). I thank Tom Baer and Hedwig Gockel for helpful comments.

References

Abel, S. M., Krever, E. M., & Alberti, P. W. (1990). Auditory detection, discrimination and speech processing in ageing, noise sensitive and hearing-impaired listeners. *Scandinavian Audiology,* 19, 43-54.

Bacon, S. P., & Viemeister, N. F. (1985). Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects. *Audiology,* 24, 117-134.

Bacon, S. P., & Gleitman, R. M. (1992). Modulation detection in subjects with relatively flat hearing losses. *Journal of Speech and Hearing Research,* 35, 642-653.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrowband carriers. *Journal of the Acoustical Society of America,* 102, 2892-2905.

Füllgrabe, C., Moore, B. C. J., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience,* 6, Article 347, 1-25.

Grose, J. H., & Mamo, S. K. (2010). Processing of temporal fine structure as a function of age. *Ear and Hearing,* 31, 755-760.

Grose, J. H., Poth, E. A., & Peters, R. W. (1994). Masking level differences for tones and speech in elderly listeners with relatively normal audiograms. *Journal of Speech and Hearing Research,* 37, 422-428.

Hawkins, D. B., & Wightman, F. L. (1980). Interaural time discrimination ability of listeners with sensori-neural hearing loss. *Audiology,* 19, 495-507.

He, N., Dubno, J. R., & Mills, J. H. (1998). Frequency and intensity discrimination measured in a maximum-likelihood procedure from young and aged normal-hearing subjects. *Journal of the Acoustical Society of America,* 103, 553-565.

Hopkins, K., & Moore, B. C. J. (2007). Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information. *Journal of the Acoustical Society of America,* 122, 1055-1068.

Hopkins, K., & Moore, B. C. J. (2010a). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America,* 127, 1595-1608.

Hopkins, K., & Moore, B. C. J. (2010b). Development of a fast method for measuring sensitivity to temporal fine structure information at low frequencies. *International Journal of Audiology,* 49, 940-946.

Hopkins, K., & Moore, B. C. J. (2011). The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise. *Journal of the Acoustical Society of America,* 130, 334-349.

Jackson, H. M., & Moore, B. C. J. (2014). The role of excitation-pattern and temporal-fine-structure cues in the discrimination of harmonic and frequency-shifted complex tones. *Journal of the Acoustical Society of America,* 135, 1356-1570.

Joris, P. X., & Yin, T. C. (1992). Responses to amplitude-modulated tones in the auditory nerve of the cat. *Journal of the Acoustical Society of America,* 91, 215-232.

King, A., Hopkins, K., & Plack, C. J. (2014). The effects of age and hearing loss on interaural phase discrimination. *Journal of the Acoustical Society of America,* 135, 342-351.

Léger, A., Heinz, M. G., Braida, L. D., & Moore, B. C. J. (2015). Sensitivity to interaural time differences in envelope and fine structure, individually and in combination. In *ARO 38th Annual Midwinter Meeting* (Abstract PS-703). Baltimore, MD: ARO.

Lutman, M. E., Gatehouse, S., & Worthington, A. G. (1991). Frequency resolution as a function of hearing threshold level and age. *Journal of the Acoustical Society of America,* 89, 320-328.

Marmel, F., Plack, C. J., Hopkins, K., Carlyon, R. P., Gockel, H. E., & Moore, B. C. J. (2015). The role of excitation-pattern cues in the detection of frequency shifts in bandpass-filtered complex tones. (in press).

Moore, B. C. J. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America,* 54, 610-619.

Moore, B. C. J. (2014). Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss. Singapore: World Scientific.

Moore, B. C. J., & Peters, R. W. (1992). Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity. *Journal of the Acoustical Society of America,* 91, 2881-2893.

Moore, B. C. J., & Glasberg, B. R. (2001). Temporal modulation transfer functions obtained using sinusoidal carriers with normally hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America,* 110, 1067-1073.

Moore, B. C. J., & Sek, A. (2009). Development of a fast method for determining sensitivity to temporal fine structure. *International Journal of Audiology,* 48, 161-171.

Moore, B. C. J., & Sek, A. (2011). Effect of level on the discrimination of harmonic and frequency-shifted complex tones at high frequencies. *Journal of the Acoustical Society of America,* 129, 3206-3212.

Moore, B. C. J., & Ernst, S. M. (2012). Frequency difference limens at high frequencies: Evidence for a transition from a temporal to a place code. *Journal of the Acoustical Society of America,* 132, 1542-1547.

Moore, B. C. J., Shailer, M. J., & Schooneveldt, G. P. (1992). Temporal modulation transfer functions for band-limited noise in subjects with cochlear hearing loss. *British Journal of Audiology,* 26, 229-237.

Moore, B. C. J., Wojtczak, M., & Vickers, D. A. (1996). Effect of loudness recruitment on the perception of amplitude modulation. *Journal of the Acoustical Society of America,* 100, 481-489.

Moore, B. C. J., Vickers, D. A., & Mehta, A. (2012). The effects of age on temporal fine structure sensitivity in monaural and binaural conditions. *International Journal of Audiology,* 51, 715-721.

Neher, T., Lunner, T., Hopkins, K., & Moore, B. C. J. (2012). Binaural temporal fine structure sensitivity, cognitive function, and spatial speech recognition of hearing-impaired listeners. *Journal of the Acoustical Society of America,* 131, 2561-2564.

Peters, R. W., & Moore, B. C. J. (1992). Auditory filters and aging: filters when auditory thresholds are normal. In Y. Cazals, L. Demany & K. Horner (Eds.). *Auditory Physiology and Perception* (pp. 179-185). Oxford: Pergamon.

Pichora-Fuller, M. K., & Schneider, B. A. (1991). Masking-level differences in the elderly: a comparison of antiphasic and time-delay dichotic conditions. *Journal of Speech and Hearing Research,* 34, 1410-1422.

Pichora-Fuller, M. K., & Schneider, B. A. (1992). The effect of interaural delay of the masker on masking-level differences in young and old subjects. *Journal of the Acoustical Society of America,* 91, 2129-2135.

Ross, B., Fujioka, T., Tremblay, K. L., & Picton, T. W. (2007). Aging in binaural hearing begins in mid-life: evidence from cortical auditory-evoked responses to changes in interaural phase. *Journal of Neuroscience,* 27, 11172-11178.

Schoof, T., & Rosen, S. (2014). The role of auditory and cognitive factors in understanding speech in noise by normal-hearing older listeners. *Frontiers in Aging Neuroscience,* 6, article 307, 1-14.

Sek, A., & Moore, B. C. J. (2012). Implementation of two tests for measuring sensitivity to temporal fine structure. *International Journal of Audiology,* 51, 58-63.

Strouse, A., Ashmead, D. H., Ohde, R. N., & Grantham, D. W. (1998). Temporal processing in the aging auditory system. *Journal of the Acoustical Society of America,* 104, 2385-2399.

Tyler, R. S., Wood, E. J., & Fernandes, M. A. (1983). Frequency resolution and discrimination of constant and dynamic tones in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America,* 74, 1190-1199.

Chapter 2

Aging Effects on Behavioural Estimates of Suppression with Short Suppressors

Erica L. Hegland, Elizabeth A. Strickland
Purdue University, West Lafayette, IN, USA

Contact author: ELH, ehegland@purdue.edu

Abstract

Auditory two-tone suppression is a nearly instantaneous reduction in the response of the basilar membrane to a tone or noise when a second tone or noise is presented simultaneously. Previous behavioural studies provide conflicting evidence on whether suppression changes with increasing age, and aging effects may depend on whether a suppressor above (high-side) or below (low-side) the signal frequency is used. Most previous studies have measured suppression using stimuli long enough to elicit the medial olivocochlear reflex (MOCR), a sound-elicited reflex that reduces cochlear amplification or gain. It has a "sluggish" onset of approximately 25 ms. There is physiological evidence that suppression may be reduced or altered by elicitation of the MOCR. In the present study, suppression was measured behaviourally in younger adults and older adults using a forward-masking paradigm with 20-ms and 70-ms maskers and suppressors. In experiment 1, gain was estimated by comparing on-frequency (2 kHz) and off-frequency (1.2 kHz) masker thresholds for a short, fixed-level 2-kHz signal. In experiment 2, the fixed-level signal was preceded by an off-frequency suppressor (1.2 or 2.4 kHz) presented simultaneously with the on-frequency masker. A suppressor level was chosen that did not produce any forward masking of the signal. Suppression was measured as the difference in on-frequency masker threshold with and without the suppressor present. The effects of age on gain and suppression estimates will be discussed.

Keywords: two-tone suppression, MOCR, aging

1 Introduction

We live in a noisy world, and understanding speech in noise is a common challenge. As people age, this task becomes increasingly difficult, and elevated thresholds alone cannot account for this increased difficulty (e.g., Pichora-Fuller et al., 1995). One explanation for this discrepancy may be attributed to the decrease in the endocochlear potential that tends to occur with increasing age (e.g., Schuknecht et al., 1974). Decreased endocochlear potential negatively affects the function of inner and outer hair cells (OHCs) (e.g., Gates et al., 2002), and disruptions in OHC function could decrease cochlear nonlinearity (e.g., Schmiedt et al., 1980).

A sensitive measure of cochlear nonlinearity is two-tone suppression, e.g., a nearly instantaneous reduction in the basilar membrane (BM) response to one tone in the presence of a second tone. There is psychoacoustic evidence that suppression measured with suppressors above the signal frequency (high-side suppression) and below the signal frequency (low-side suppression) may show different aging effects. Several of these psychoacoustic studies have

used a band-widening technique in a forward-masking paradigm to measure suppression (Dubno & Ahlstrom, 2001a, b; Gifford & Bacon, 2005). Narrow bandpass noise centered at the signal frequency was widened by adding noise bands above, below, or above and below the signal frequency. Suppression was measured as the difference in signal threshold between the narrowest and widest noisebands. Dubno and Ahlstrom (2001a, b) found less suppression with more noise above rather than below the signal frequency, and older adults with normal hearing had less suppression than the younger adults. Older adults had minimal low-side suppression and absent high-side suppression (Dubno & Ahlstrom, 2001a). In contrast, using the band-widening technique with noise added below the signal, Gifford and Bacon (2005) found no effect of age on suppression. Using tonal maskers and suppressors in a forward-masking paradigm, Sommers & Gehr (2010) found reduced or absent high-side suppression with increased age.

It is possible that previous psychoacoustic studies of suppression underestimated the amount of suppression by using maskers that were long enough to elicit the medial olivocochlear reflex (MOCR). The MOCR is a sound-elicited efferent feedback loop that reduces OHC amplification or gain (Cooper & Guinan, 2006). Gain starts to be affected approximately 25 ms after elicitor onset. Evidence from physiological studies in animals suggests that elicitation of the MOCR may decrease or modify suppression(Winslow & Sachs, 1988; Kawase et al., 1993). Previous studies of aging effects on suppression used masker and suppressor durations of 200-500 ms (Dubno & Ahlstrom, 2001a, b; Gifford & Bacon, 2005; Sommers & Gehr, 1998, 2010). These durations were long enough to elicit MOCR-induced gain reductions and possibly reduce suppression estimates.

The goal of this study is to investigate the effect of age on estimates of two-tone suppression using shorter stimuli. Several prior studies of aging effects on suppression have used masker levels that had been previously shown to produce large amounts of suppression in young adults (Dubno & Ahlstrom, 2001a, 2001b; Gifford & Bacon, 2005; Sommers & Gehr, 2010). In the present study, suppressor levels were chosen that did not produce forward masking of the signal, which can be presumed to produce minimal spread of excitation at the signal frequency place (Moore & Vickers, 1997). If simultaneous masking is produced by spread of excitation and suppression, using a suppressor that does not produce forward masking should result in a measure of suppression rather than a mixture of suppression and excitation.

2 Methods

2.1 Participants

Participants included five younger adults between 18 and 30 years old (mean = 21.4; 3 female) and five older adults between 62 and 70 years old (mean = 65.4; 4 female). The average threshold for younger adults at 2 kHz was 5 dB HL (SD 3.54), and the average threshold for older adults at 2 kHz was 15 dB HL (SD 9.35). The Montreal Cognitive Assessment (MoCA; used with permission, copyright Z. Nasreddine) was administered, and all results were within the normal range (≥ 26/30).

2.2 Stimuli

The signal was a 10-ms, 2-kHz sinusoid with 5-ms onset and offset ramps and no steady state portion. This signal frequency has been used in previous studies on aging effects on suppression (Sommers & Gehr, 1998, 2010; Dubno & Ahlstrom, 2001a, 2001b; Gifford & Bacon, 2005). The signal was presented at 10 dB SL and/or 50 dB SPL for each participant, and the masker level was adaptively varied to find threshold. Masker durations were 20 (short) and 70 ms (long). The 20-ms masker should be too short to elicit the MOCR, while the 70-ms masker should be long enough to elicit full gain reduction from MOCR elicitation (Roverud & Strickland, 2010). For experiment 1, masker thresholds for the short and long on-frequency (2-kHz) and off-frequency (1.2 and 2.4 kHz) maskers were measured (Figure 1). At the signal frequency place, the BM

response to the 1.2-kHz masker should be approximately linear, even with MOCR stimulation (Cooper & Guinan, 2006). The 2.4-kHz masker frequency was chosen because it has produced substantial suppression of a 2-kHz masker in previous studies of young adults (Shannon, 1976; Sommers & Gehr, 2010). Gain was estimated as the difference in threshold between the on-frequency masker and the 1.2-kHz masker (Yasin et al., 2014). For experiment 2, suppression was measured by presenting a short or long, 1.2- or 2.4-kHz suppressor simultaneously with the on-frequency masker and just prior to presentation of the 10-ms, 2-kHz signal. This will be referred to as the "combined masker." The masker and suppressor durations were always equal. The on-frequency masker level was adaptively varied to find threshold.

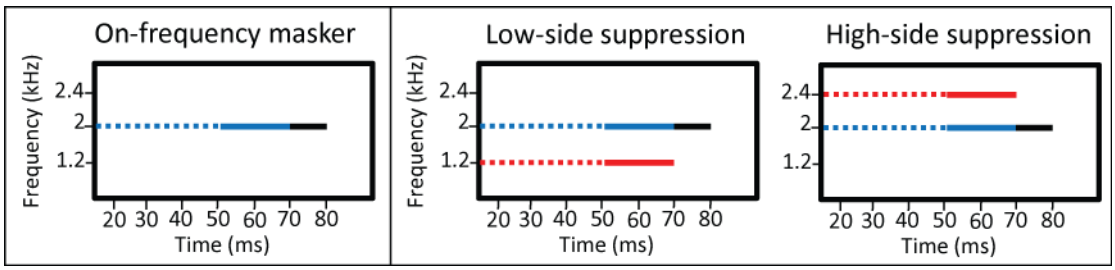


Figure 1. Schematic of stimuli frequencies (kHz) and durations (ms). Masker durations were 20 ms (solid lines) and 70 ms (solid + dashed lines). Maskers were on-frequency (blue, 2 kHz) and off-frequency (red, 1.2 and 2.4 kHz).

All stimuli were digitally produced with MATLAB software (2011a, The Math Works, Natick, MA). They were output via a Lynx II XLR sound card to a headphone buffer (TDT HB6) and presented to the right ear or the better ear of each participant through an ER-2 insert earphone (Etymotic Research, Inc, Elk Grove Village, IL).

2.3 Procedure

Participants sat in a sound-attenuating booth and listed to stimuli presented to their right ear or better ear. Masker threshold for a fixed signal level was measured using a three-interval forced-choice adaptive procedure. Participants were instructed to use a computer mouse or keyboard to choose the interval that sounded different from the others. The masker level was increased following one incorrect response and decreased following two correct responses. Following this procedure, an estimate of 70% correct on the psychometric function is obtained (Levitt, 1971). Masker threshold for a given trial was calculated as an average of the last eight of the 12 total reversals. Thresholds with a standard deviation of 5 dB or greater were not included in the final average. The order of the conditions was counterbalanced across the participants.

3 Results

3.1 Experiment 1

Thresholds for the short and long on-frequency (2 kHz) and off-frequency maskers (1.2 and 2.4 kHz) were obtained for signal levels fixed at 10 dB above each participant's threshold for the 2-kHz, 10-ms tone. Because the absolute signal level differed between younger and older participants, masker thresholds were also obtained for a 50-dB SPL signal, which was 10-dB SL for some older participants.

Estimates of cochlear gain were calculated as the difference in masker threshold between the 2-kHz and 1.2-kHz short maskers (Yasin et al., 2014). Gain was determined separately for each signal level. A one-way ANOVA was conducted to compare the effect of age group on gain estimates at each signal level. There was a significant effect of age group on gain estimates at the 10-dB SL signal level [$F(1,8) = 7.03$, $p = 0.029$] and the 50-dB SPL signal level [$F(1,7) = 6.86$, $p =$

0.034]. Average amounts of gain are shown in Figure 2. These results suggest that gain estimates were significantly decreased in older adults compared to younger adults. Correlation coefficients were computed between auditory threshold at 2 kHz and gain estimates at both signal levels, and neither of the correlations reached statistical significance.

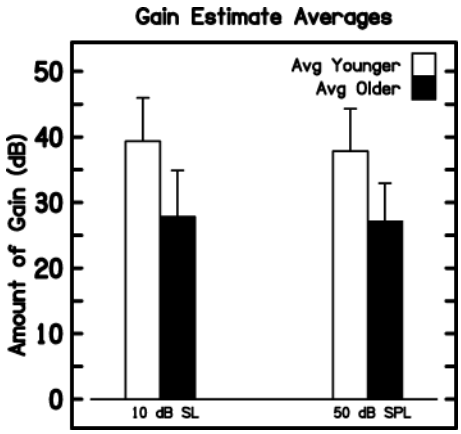


Figure 2. Average amounts of gain for older and younger listeners at the 10-dB SL and 50-dB SPL signal levels.

3.2 Experiment 2

Suppression was measured in a forward-masking paradigm (see Figure 1). Suppressor levels were chosen that were below the thresholds obtained in Experiment 1 to ensure that the suppressor did not produce any forward masking of the signal. The suppressor was an average of 4 dB below the off-frequency masker threshold obtained in Experiment 1 [*SD* = 1.60]. For younger adults, suppressors were presented at an average of 74.50 dB [*SD* = 5.22] for low-side suppression and 74.23 dB [*SD* = 4.00] for high-side suppression. For older adults, suppressors were presented at an average of 67.60 dB [*SD* = 2.12] for low-side suppression and 75.20 dB [*SD* = 3.68] for high-side suppression.

Average amounts of low-side and high-side suppression for younger adults (open bars) and older adults (filled bars) at the 10-dB SL and 50-dB SL signal levels are shown in Figure 3. With the short combined masker, all participants showed some low-side suppression, and all participants except for one older adult showed high-side suppression. With the long combined masker, 2/5 older adults had no measurable low-side suppression, and 3/5 older adults had no measurable high-side suppression. A one-way ANOVA was conducted to compare the effect of age group on suppression estimates. There was a significant effect of age group on suppression estimates at the $p < .05$ level for all conditions tested. There was a significant effect of age group for low-side suppression with a short combined masker [10-dB SL signal: $F(1,8) = 5.53$, $p = 0.047$; 50-dB SPL signal: $F(1,8) = 11.09$, $p = 0.010$] and with a long combined masker [10-dB SL signal: $F(1,8) = 10.57$, $p = 0.012$; 50-dB SPL signal: $F(1,8) = 12.94$, $p = 0.007$]. There was a significant effect of age group for high-side suppression with a short combined masker [10-dB SL signal: $F(1,8) = 8.66$, $p = 0.019$; 50-dB SPL signal: $F(1,7) = 10.37$, $p = 0.015$] and with a long combined masker [10-dB SL signal: $F(1,8) = 7.55$, $p = 0.025$; 50-dB SL signal: $F(1,7) = 7.99$, $p = 0.026$]. This suggests that both high-side and low-side suppression estimates decreased with age.

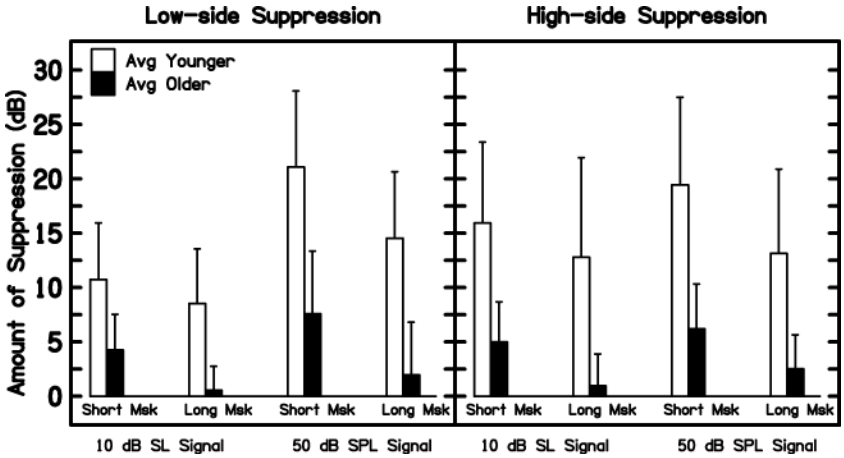


Figure 3. Average suppression estimates for younger adults (open bars) and older adults (filled bars). Results are shown for the 1.2 kHz suppressor (low-side suppression), 2.4 kHz suppressor (high-side suppression), 20-ms combined masker (short msk), 70-ms combined masker (long msk), 10-dB SL signal level, and 50-dB SPL signal level.

A paired-samples t-test was conducted to compare suppression estimates obtained with short and long combined maskers. In all suppression conditions tested, there were significant differences at the $p < .05$ level between suppression estimates obtained with short and long combined maskers. There was a significant difference between low-side suppression estimates obtained with short [M = 7.49, SD = 5.34] and long combined maskers [M = 4.54, SD = 5.57] at the 10-dB SL signal level [$t(9) = 2.94$, $p = 0.017$] and between short [M = 14.33, SD = 9.33] and long combined maskers [M = 8.25, SD = 8.42] at the 50-dB SPL signal level [$t(9) = 3.41$, $p = 0.008$]. There was a significant difference between high-side suppression estimates obtained with short [M = 10.46, SD = 7.99] and long combined maskers [M = 6.86, SD = 8.95] at the 10-dB SL signal level [$t(9) = 3.83$, $p = 0.004$] and between short [M = 12.08, SD = 9.03] and long combined maskers [M = 7.24, SD = 7.66] at the 50-dB SPL signal level [$t(9) = 4.56$, $p = 0.002$]. This suggests that suppression estimates were significantly greater when measured with a short combined masker.

It is possible that the decreased suppression in the older adults was due to their increased auditory thresholds. To determine the effect of auditory threshold on suppression estimates, correlation coefficients were computed between threshold in dB HL and suppression estimates from each of the eight conditions tested. Out of the eight suppression conditions, threshold at 2 kHz was negatively correlated with low-side suppression at the 50-dB SPL signal level measured with the short combined masker [$r(10) = -0.68$, $p = 0.031$] and with the long combined masker [$r(10) = -0.67$, $p = 0.33$]. This suggests that those with higher auditory thresholds had smaller low-side suppression estimates at the 50-dB SPL signal level. Otherwise, auditory threshold was not significantly correlated with either suppression or gain estimates.

4 Discussion

In the present experiment, several important differences were seen between younger adults and older adults. In Experiment 1, older adults were found to have less gain than younger adults, measured as the difference in on-frequency and off-frequency masker threshold for a fixed signal level (Yasin et al., 2014). The off-frequency, 1.2-kHz masker is understood to have a linear BM response at the signal frequency place, and the on-frequency, 2-kHz masker and signal have a compressive BM response at the signal frequency place (Oxenham & Plack, 2000). Comparing these two conditions gives an estimate of cochlear gain around the signal frequency place. It is possible that the decrease in gain for the older adults was due to their slightly poorer auditory thresholds. However, correlation coefficients between auditory threshold and gain estimates did

not reach significance. This measure of gain may reflect decreases in cochlear function that occur prior to elevations in auditory thresholds. Decreased endocochlear potential may have less of an effect on thresholds closer to the apex of the cochlea and a larger effect in the base (Mills et al., 2006) while still impacting cochlear nonlinearity.

Estimates of high-side and low-side suppression were significantly smaller for older adults than for younger adults. Most previous studies of aging effects on suppression have found decreased or absent suppression for older adults (Sommers & Gehr, 1998, 2010; Dubno & Ahlstrom, 2001a, b). In the present study, suppressor levels were chosen to not produce forward masking and to result in the greatest measure of suppression. Also, with the long combined masker, some older adults did not show high-side suppression, and some did not show low-side suppression. By using short combined maskers with individually-selected suppressor levels, accurate measurements of suppression can be achieved.

Both younger and older adults in the present study showed significantly more suppression measured with short combined maskers than with longer combined maskers. Several of the older participants, who had no measurable suppression with the longer combined masker, had a small amount of suppression measured with the shorter combined masker. When suppression is measured with combined maskers long enough to elicit MOCR-induced gain reduction, suppression may be underestimated. However, even when measured with the shorter combined maskers, older adults had less suppression than younger adults. Evidence from animal studies suggests that gain reduction from the MOCR and suppression interacts, resulting in decreased suppression (Winslow & Sachs, 1988; Kawase et al., 1993). The results of the present study support this idea.

Dubno and Ahlstrom (2001a) found a negative correlation between auditory thresholds and suppression estimates. In the present study, there was a negative correlation between auditory thresholds and suppression estimates for low-side suppression at the 50-dB SPL signal level. The correlation was not significant for low-side suppression at the 10-dB SL signal level or for high-side suppression. The suppression estimates in Dubno and Ahlstrom (2001a) were measured between the narrowest and widest bandwidth maskers. This measure would include elements of both high-side and low-side suppression. It is possible that auditory threshold is more strongly correlated with low-side suppression.

The present study found strong evidence for age-related decreases in both high-side and low-side two-tone suppression and in estimates of cochlear gain. These findings are consistent with decreasing endocochlear potential that occurs with increasing age. The results also support the hypothesis that that suppression estimates may be underestimated when measured with long combined maskers, which would be consistent with a reduction in suppression due to the MOCR.

Acknowledgments

This research was supported by NIH(NIDCD) T32 DC000030-21 and NIH(NIDCD) F31 DC014395-01A1 (first author), and NIH(NIDCD) R01 DC008327 (second author).

References

Cooper, N. P., & Guinan, Jr., J. J. (2006). Efferent-mediated control of basilar membrane motion. *Journal of Physiology*, *576.1*, 49–54.

Dubno, J. R., & Ahlstrom, J. B. (2001a). Psychophysical suppression effects for tonal and speech signals. *Journal of the Acoustical Society of America*, *110*(4), 2108–2119.

Dubno, J. R., & Ahlstrom, J. B. (2001b). Psychophysical suppression measured with bandlimited noise extended below and/or above the signal: Effects of age and hearing loss. *Journal of the Acoustical Society of America*, *110*(2), 1058–1066.

Gates, G. A., Mills, D., Nam, B., D'Agostino, R., & Rubel, E. W. (2002). Effects of age on the distortion product otoacoustic emission growth functions. *Hearing Research*, *163*(1-2), 53–60.

Gifford, R. H., & Bacon, S. P. (2005). Psychophysical estimates of nonlinear cochlear processing in younger and older listeners. *Journal of the Acoustical Society of America*, *118*(6), 3823–3833.

Kawase, T., Delgutte, B., & Liberman, M. C. (1993). Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones. *Journal of Neurophysiology*, *70*(6), 2533–2549.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*(2), 467–477.

Mills, J. H., Schmiedt, R. A., Schulte, B. A., & Dubno, J. R. (2006). Age-related hearing loss: A loss of voltage, not hair cells. *Seminars in Hearing*, *27*(4), 228–236.

Moore, B. C. J., & Vickers, D. A. (1997). The role of spread of excitation and suppression in simultaneous masking. *Journal of the Acoustical Society of America*, *102*(4), 2284–2290.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Roverud, E., & Strickland, E. A. (2010). The time course of cochlear gain reduction measured using a more efficient psychophysical technique. *Journal of the Acoustical Society of America*, *128*(3), 1203–1214.

Schmiedt, R. A., Zwislocki, J. J., & Hamernik, R. P. (1980). Effects of hair cell lesions on responses of cochlear nerve fibers. I. Lesions, tuning curves, two-tone inhibition, and responses to trapezoidal-wave patterns. *Journal of Neurophysiology*, *43*(5), 1367–1389.

Schuknecht, H. F., Watanuki, K., Takahashi, T., Belal, Jr., A. A., Kimura, R. S., Jones, D. D., & Ota, C. Y. (1974). Atrophy of the stria vascularis, a common cause for hearing loss. *The Laryngoscope*, *84*(10), 1777–1821.

Shannon, R. V. (1976). Two-tone unmasking and suppression in a forward-masking situation. *Journal of the Acoustical Society of America*, *59*(6), 1460–1470.

Sommers, M. S., & Gehr, S. E. (1998). Auditory suppression and frequency selectivity in older and younger adults. *Journal of the Acoustical Society of America*, *103*(2), 1067–1074.

Sommers, M. S., & Gehr, S. E. (2010). Two-tone auditory suppression in younger and older normal-hearing adults and its relationship to speech perception in noise. *Hearing Research*, *264*, 56–62.

Winslow, R. L., & Sachs, M. B. (1988). Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle. *Hearing Research*, *35*, 165–190.

Yasin, I., Drga, V., & Plack, C. J. (2014). Effect of human auditory efferent feedback on cochlear gain and compression. *The Journal of Neuroscience*, *34*(46), 15319–15326.

Chapter 3

Contributions of Coding Efficiency of Temporal-Structure and Level Information to the Lateralization Performance in Young and Early-Elderly Listeners

Atsushi Ochi, Tatsuya Yamasoba

Department of Otolaryngology, Faculty of Medicine, University of Tokyo, Tokyo, Japan

Shigeto Furukawa

Human Information Science Laboratory, NTT Communication Science Laboratories, NTT Corporation, Atsugi, Japan

Contact author: SF, furukawa.shigeto@lab.ntt.co.jp

Abstract

The performance of lateralization task based on interaural time or level difference (ITD or ILD) often varies among listeners. This study examined the extent to which this inter-listener variation could be accounted for by the coding efficiency of temporal-structure or level information below the stage of interaural interaction. Young listeners (20s to 30s) and early-elderly (60s) listeners with or without mild hearing loss were tested. The *ITD*, *ILD*, *TIME*, and *LEVEL* tasks were intended to measure sensitivities to ITD, ILD, the temporal structure of the stimulus encoded by the neural phase locking, and stimulus level, respectively. The performances of the *ITD* and *ILD* tasks were not significantly different between the age groups, while the elderly listeners exhibited significantly poorer performance in the *TIME* task (and in the *LEVEL* with high frequency stimulus only) than the young listeners. Significant correlations were found between thresholds for the *ILD* and *LEVEL* tasks with low- and high-frequency stimuli and for the *ITD* and *TIME* tasks for the high-frequency stimulus, implying peripheral coding efficiency as a major factor that determines lateralization performance. However, we failed to find a correlation between the *ITD* and *TIME* tasks for the low-frequency stimulus, despite a large range of threshold values of the *TIME* task. This implies that in a low frequency region, the peripheral coding efficiency of stimulus temporal structure is a relatively minor factor for the ITD-based lateralization performance.

Keywords: interaural time difference, interaural level difference, temporal structure, intensity coding, inter-individual variation, aging

1 Introduction

Interaural time and level differences (ITDs and ILDs) are the major cues for horizontal sound localization. Sensitivities to ITD and ILD, evaluated by lateralization tasks, often vary markedly among listeners. Lateralization performance based on ITD and ILD should reflect not only the listener's ability to compare time and level information, respectively, between ears, but also the

efficiency of encoding information about temporal structure and intensity of stimuli at stages below binaural interactions in auditory processing. Our earlier study attempted to evaluate the relative contributions of these processing stages to the inter-listener variability in lateralization performance, by comparing individual listeners' monaural sensitivities to the temporal structure and intensity of a sound stimulus with their ITD and ILD sensitivities (Ochi et al., 2014). The results showed significant correlation of ILD discrimination thresholds with thresholds for monaural level-increment detection task. This suggests that that the inter-individual differences in ILD sensitivity could be accounted for (partially) by the level coding efficiency at stages before binaural interaction. Similarly, ITD discrimination thresholds were found to correlate with the listeners' sensitivities to the temporal structure of monaural stimuli, when the stimuli were in high frequency range (around 4000 Hz). However, we failed to find a positive correlation for stimuli in lower frequency range (around 1100 Hz).

The present study extended our earlier study (Ochi et al., 2014) by incorporating elderly listeners under essentially the same experimental settings. We first examined the effects of age on the performances of individual tasks. We then analysed correlations, as in the earlier study, between the task performances. It has been reported that the sensitivities to temporal structure and intensity of stimuli decline with age (e.g., Hopkins & Moore, 2011). A population including young and elderly listeners would therefore exhibit a large variability of thresholds in the monaural tasks, which would lead to improved sensitivities of the correlation analysis and provide further insights as to the roles of monaural processing in ITD or ILD discrimination. Supplemental data have been obtained also to evaluate underlying mechanisms for the monaural tasks.

As in the earlier study, we measured listeners' performances in four basic tasks, namely *ITD*, *ILD*, *TIME*, and *LEVEL* tasks, which would reflect sensitivities to ITD, ILD, temporal structure, and level change of stimuli, respectively. Low- and high-frequency stimuli were tested, which were centred around 1100 and 4000 Hz, respectively. Supplementary experiments measured frequency resolution (*FRES* task), frequency discrimination limen (*FDISC* task), and (for low-frequency stimulus only) discrimination threshold of Huggins pitch (*HUGGINS* task; Cramer & Huggins, 1958).

2 Methods

2.1 Listeners

Forty-three adults participated in the experiment. All gave written informed consent, which was approved by the Ethics Committee of NTT Communication Science Laboratories. Those included 22 normal-hearing young listeners (referred to as the YNH group; 10 males and 12 females; 19-43 years old, mean 32.0) and 21 elderly listeners (11 males and 10 females; 60-70 years old, mean 63.0). The YNH listeners have been represented in the earlier study (Ochi et al., 2014). In the analyses, the elderly listeners were further divided into normal-hearing (referred to as ENH) and hearing-impaired (EHI) groups. Listeners with audiometric thresholds of <30 dB HL at all the frequencies between 125 and 4000 Hz in both ears were classified as normal-hearing, otherwise as (mildly) hearing-impaired. For the *FDISC* and *HIGGINS* tasks, A subset of YNH listeners (N = 12) and all the elderly listeners participated. Table 1 summarizes the means and standard deviations of hearing levels obtained by pure-tone audiometry.

Table 1. Means and standard deviations of hearing levels for the three listener groups. Columns represent, from left to right, averages across all the frequencies, 1000-Hz tone, and 4000-Hz tone, respectively.

	Hearing Level (dB) mean ± standard deviation		
	125-4000 Hz	1000 Hz	4000 Hz
YNH (N=22)	8.1 ± 6.5	5.5 ± 5.6	3.1 ± 6.5
ENH (N=12)	13.0 ± 6.1	10.2 ± 7.0	11.7 ± 6.2
EHI (N=9)	21.1 ± 13.4	11.9 ± 7.1	35.3 ± 18.8

2.2 Stimuli

Stimuli were presented to the listener through headphones. Except for the binaural tasks (i.e., *ITD*, *ILD*, and *HUGGINS* tasks), the stimuli were presented to the right ear.

The main four tasks (namely, *ITD*, *ILD*, *TIME*, and *LEVEL* tasks) employed two types of stimuli, referred to as the low- and high-frequency stimuli, which were identical to those used in our earlier study (Ochi et al., 2014), and thus are described briefly here. The stimuli were designed to assess the listener's ability to use information based on neural phase-locking to the stimulus temporal structure, respectively, in the *ITD* and *TIME* tasks. Essentially the same stimuli were used also in the *ILD* and *LEVEL* tasks. The low-frequency stimulus was a spectrally shaped multicomponent complex (SSMC), which was a harmonic complex with a fundamental frequency (F_0) of 100 Hz. The spectral envelope had a flat passband and sloping edges (5xF0 centered at 1100 Hz).The overall level of the complex was 54 dB SPL. Threshold equalizing noise, extending from 125 to 15000 Hz, was added. The high-frequency stimulus was a "transposed stimulus," which was a 4-kHz tone carrier amplitude-modulated with a half-wave rectified 125-Hz sinusoid (Bernstein & Trahiotis, 2002). It is considered that the auditory-nerve firing is phase locked to the modulator waveform, which provides the cue for judging the ITD and modulation rate of the stimulus. The overall level of the transposed stimulus was set to 65 dB SPL. A continuous, low-pass filtered Gaussian noise was added to prevent the listener from using any information at low spectral frequencies (e.g., combination tones).

Stimuli used for supplementary tasks (namely, *FRES* and *FDISC* tasks) involved tone-burst signals at frequencies of 1100 and 4000 Hz. When the two tasks are concerned, the low- and high-frequency stimuli refer to the 1100- and 4000-Hz tones, respectively. The frequency band of interest in the *HUGGINS* task (another supplementary task) was centred at 1100 Hz. Other details about the stimuli for the FRES, FDISC, and HUGGINS tasks are described in the subsequent section.

2.3 Procedures

General procedure

A two-interval two-alternative forced-choice (2I-2AFC) method was used to measure the listener's sensitivities to stimulus parameters. Feedback was given to indicate the correct answer after each response. The two-down/one-up adaptive tracking method was used to estimate discrimination thresholds.

Task specific procedures

ITD task — In a 2I-2AFC trial, stimuli in the two intervals had ITDs of +ΔITD/2 and -ΔITD/2 μs, respectively. Each stimulus was 400-ms long, including 100-ms raised-cosine onset and offset

ramps, which were synchronized between the two ears. The listeners were required to indicate the direction of the ITD change between the two intervals on the basis of the laterality of sound images.

ILD task — Similarly to the *ITD* task, stimuli in the two intervals had ILDs of +ΔILD/2 and -ΔILD/2 dB, respectively. Each stimulus was 400-ms long, including 20-ms raised-cosine onset and offset ramps. The listeners were required to indicate the direction of the ILD change between the two intervals on the basis of the laterality of sound images.

TIME task — For the low-frequency stimulus, the listeners were required to detect a common upward frequency shift (Δf Hz) imposed on the individual components of the SSMC with the spectral envelope remaining unchanged (Moore & Sek, 2009). It was assumed that the listeners based their judgments on pitch changes, reflecting the temporal fine structure encoded as the pattern of neural phase locking. The "signal" and "non-signal" intervals in the 2I-2AFC method contained RSRS and RRRR sequences, respectively, where R indicates original SSMC and S indicates a frequency-shifted SSMC. For the high-frequency stimulus, the listener's task was to discriminate the modulation frequencies of the transposed stimuli between f_m (=125 Hz) and f_m + Δf Hz, referred to as R and S, respectively. Each R and S had a duration of 100 ms, including 20-ms raised-cosine ramps. The threshold was expressed as $\Delta f/f_0$ or $\Delta f/f_m$ for the low- or high-frequency stimuli, respectively. When adaptive tracking failed to converge within this limit, trials with a shift of $0.5F_0$ were repeated 30 times. In that case, the proportion of correct trials was converted to *d'*, and then the "threshold" was derived on the assumption that *d'* is proportional to the frequency shift (Moore & Sek, 2007).

LEVEL task — In a 2I-2AFC trial, the listeners were required to indicate an interval containing a 400-ms-long SSMC or a transposed stimulus whose central 200-ms portion (including 20-ms raised-cosine ramps) was incremented in level by ΔL dB, while the other non-signal interval contained an original SSMC or a transposed stimulus.

FRES task — The notched-noise masking method (Patterson et al., 1982) was employed to evaluate frequency selectivity. The signals were pure-tone busts centred at 1100 or 4000 Hz, and the maskers were notched noises with varying notch width (0, 0.05, 0.1, 0.2, 0.3, and 0.4 relative to the signal frequency). The spectrum level within the passband was 40 dB SPL. A rounded-exponential filter (Patterson et al., 1982) was fitted to the experimental data using a least-square fit. The equivalent rectangular bandwidth (*ERB*) was then derived from the parameters of the fitted filter.

FDISC task — Frequency difference limens were measured with pure-tone bursts centred at 1100 and 4000 Hz. Similarly to the *TIME* task, the sequence of RRRR and RSRS was presented, R and S representing tone bursts with frequencies of fc and f_c+Δf Hz, respectively (fc = 1100 or 4000 Hz). Each tone burst had 200ms of duration with 20ms onset-offset ramps.

HUGGINS task — Bandpass-filtered noises (passband: 250-4000 Hz) with a duration of 200 ms were used as stimuli. The noise was diotic except for a narrow frequency band centred around 1100 Hz (fc) with a 18% width around the centre frequency, on which interaural phase transition was imposed. This stimulus elicits a sensation of pitch corresponding to f_c (Cramer & Huggins, 1958). Similarly to the *TIME* task, the sequence of RRRR and RSRS was presented, R and S representing tone bursts with frequencies of fc and f_c+Δf Hz, respectively. The discrimination threshold was expressed as $\Delta f/f_c$.

2.4 Data Analyses

MATLAB with Statistical Toolbox was used for statistical analyses of the data. For the *ITD*, *TIME*, *LEVEL*, *FDISC*, and *HUGGINS* tasks, the analyses were performed with log-transformed threshold data.

3 Results

Left four columns of panels of Figure 1 compare the performance of the four basic tasks (*ITD*, *ILD*, *TIME*, and *LEVEL*) between listener groups. For the low-frequency stimulus, one-way analysis of variance indicated a statistically significant effect of listener group in the *ITD* task only. Subsequent pair-wise comparisons indicated higher thresholds for the ENH and EHI listeners than for the YNH listeners. For the high-frequency stimulus, significant effect of listener group was found in the *ITD*, *TIME*, and *LEVEL* tasks. These results reflect higher thresholds for *EHI* than for *ENH* (*ITD* task); for ENH and EHI than for YNH (*TIME* task); and for ENH than for YNH (*LEVEL* task). Listener-group effect was not significant for the ILD task in either stimulus type. The figure also shows the data obtained by supplementary experiments (*FRES*, *FDISC*, and *HUGGINS* tasks). Significant listener-group effects were found in the *HUGGINS* and (for the high-frequency stimulus only) *FDISC* tasks.

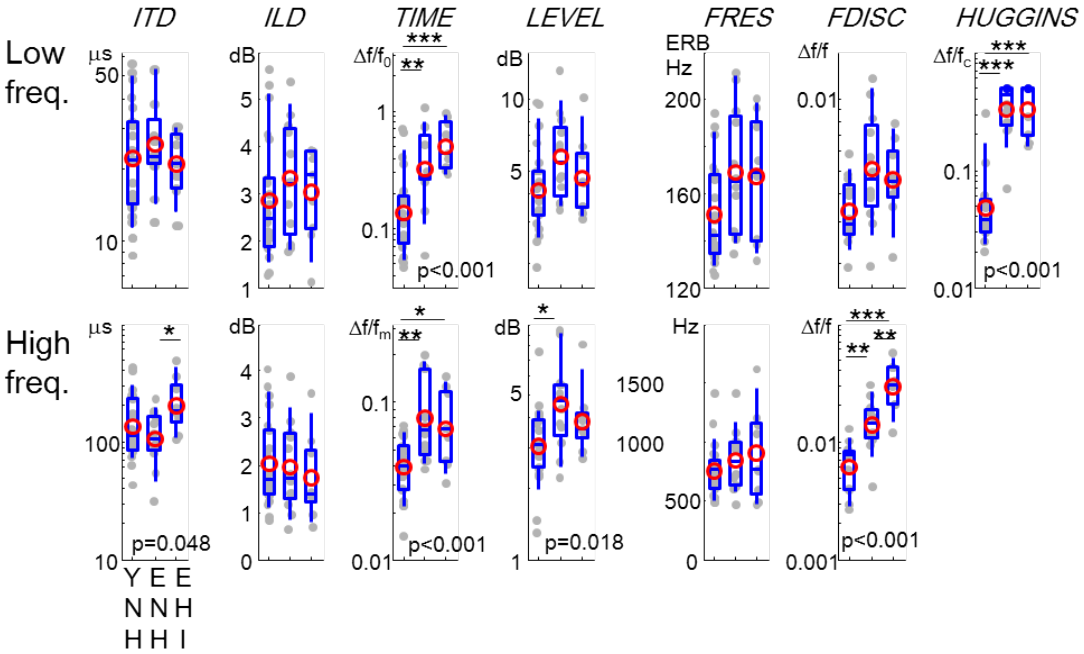


Figure 1. Comparisons of thresholds among listener groups. Each column of panels represents one task, as labelled. The three columns of data within a panel correspond to, from left to right, YNH, ENH, and EHI, respectively. The 10th, 25th, 50th, 75th, and 90th percentiles are shown by the box and whisker plot. The red circle shows the mean, and the grey dots represent individual listeners' data. The task for which a significant effect of listener group was revealed by one-way analysis of variance was indicated by the *p*-values. Asterisks indicate group pairs for which a significant difference was indicated by a post hoc tast with Tukey's honestly significant difference criterion (*: p<0,05; **: p<0.01; ***: p<0.001).

The thresholds of individual listeners are compared between pairs of tasks in Figure 2. The results showed statistically significant positive correlations for the *ITD-ILD* and *ILD-LEVEL* pairs. The partial correlation coefficients by controlling the effect of age (YNH versus ENH/EHI) were also significant. A significant partial correlation coefficient was found also for the *ITD-TIME* pair with high-frequency stimulus.

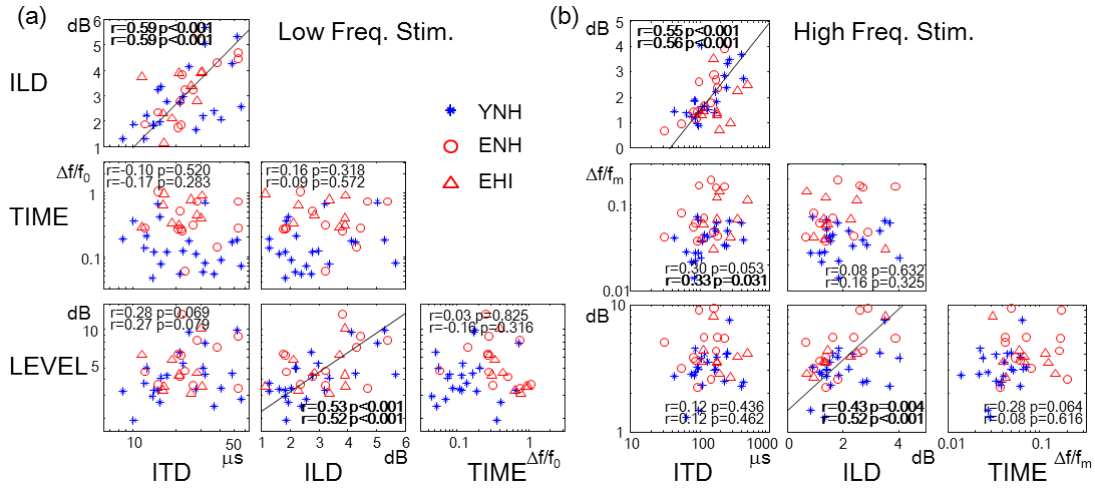


Figure 2. Comparisons of individual listeners' thresholds between tasks. (a) and (b) represent low- and high-frequency stimuli, respectively. Pearson's correlation coefficients and *p*-values are shown in the first line of numbers within a panel. Significant correlations were marked as bold characters, and orthogonal-fitting lines. The second line indicates the correlation coefficients and p-values in the when the effect of age was partialed out.

4 Discussion

Elderly listeners (ENH and EHI) showed higher thresholds than young listeners (YNH) in the *TIME* task (with both low- and high-frequency stimuli) and the *HUGGINS* task. This confirms earlier findings that the sensitivity to temporal structure declines without accompanying elevated audiometric thresholds (e.g., Hopkins & Moore, 2011). Thresholds for the *LEVEL* task also tended to be higher in the elderly than the young, consistently with previous finding (e.g., He et al., 1998). Despite these consistent declines of the performances of the monaural tasks with age, the present study failed to find significant age effect in the lateralization tasks (*ITD* and *ILD*).

The thresholds for the *ILD* and *LEVEL* tasks correlated for both low- and high-frequency stimuli. This confirms the results of our earlier study (Ochi et al., 2014), suggesting that the efficiency of level coding in the auditory periphery is a major factor that accounts for inter-individual variation of ILD sensitivity. The *ITD* task showed a correlation with the *TIME* task for the high-frequency stimulus, when the factors of age and hearing-impairments were controlled. This again is consistent with the finding of the earlier study (Ochi et al., 2014), suggesting that a listener's ITD sensitivity is accounted well for by the listener's sensitivity to temporal (envelope) structure.

Despite a relatively large number of participants and greater range of threshold values of the *TIME* task, however, we failed to find a correlation between the *ITD* and *TIME* tasks for the low-frequency stimulus. The thresholds for the *HUGGINS* task showed a significant positive correlation with those for the *TIME* task, but not with those for the *ITD* task. This suggests that the performances of the *TIME* and *HUGGINS* tasks capture inter-individual variation of the efficiency of temporal-structure processing, but that of the *ITD* task is determined primarily by other factors. The *HUGGINS* and *TIME* tasks are similar in terms that both required interaural comparison of temporal-structure information, but differs in the perceptual domain on which the listeners were expected to respond (i.e., pitch versus laterality).

It should be noted, however, that the positive correlation found for the *TIME* and *HUGGINS* tasks was due predominantly to the consistent age effect in the both tasks (see Figure 1). When the effects of age were partialed out, the correlation coefficient was not significantly from zero, implying that the positive correlation was due to underlying non-temporal factors that are

sensitive to aging. A candidate of such a factor is frequency selectivity. It has been argued that (intermediately) resolved frequency components of the SSMC could contribute to the performance of the *TIME* task (Micheyl et al., 2010). Peripheral frequency resolution could influence the representation of interaural correlation across a frequency axis, which is a basis for the Huggins pitch. Indeed, ERB was significantly correlated with thresholds for the *TIME* and *HUGGINS* tasks, the effects of age being not partialed out.

Acknowledgments

This study was supported by an internal basic research fund of NTT Corporation.

References

Bernstein, L.R., and Trahiotis, C. (2002). Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli". *Journal of the Acoustical Society of America*, 112, 1026-1036.

Cramer, E. M., & Huggins, W. H. (1958). Creation of Pitch through Binaural Interaction. *Journal of the Acoustical Society of America, 30*(5), 413-417.

He, N., Dubno, J. R., & Mills, J. H. (1998). Frequency and intensity discrimination measured in a maximum-likelihood procedure from young and aged normal-hearing subjects. *Journal of the Acoustical Society of America, 103*(1), 553-565.

Hopkins, K., & Moore, B. C. (2011). The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise.. *Journal of the Acoustical Society of America, 130*(1), 334-349.

Micheyl, C., Dai, H., & Oxenham, A. J. (2010). On the possible influence of spectral- and temporal-envelope cues in tests of sensitivity to temporal fine structure. *Journal of the Acoustical Society of America, 127*(3), 1809-1810.

Moore, B.C., & Sek, A. (2009). Development of a fast method for determining sensitivity to temporal fine structure. International *Journal of Audiology*, 48, 161-171.

Ochi, A., Yamasoba, T., & Furukawa, S. (2014). Factors that account for inter-individual variability of lateralization performance revealed by correlations of performance among multiple psychoacoustical tasks. *Frontiers in Neuroscience, 8*. doi: 10.3389/fnins.2014.00027

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., & Milroy, R. (1982). The deterioration of hearing with age: frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journal of the Acoustical Society of America, 72*(6), 1788-1803.

Chapter 4

Investigating the Role of Working Memory in Speech-in-Noise Identification for Listeners with Normal Hearing

Christian Füllgrabe
MRC Institute of Hearing Research, Nottingham, UK

Stuart Rosen
UCL Speech, Hearing & Phonetic Sciences, London, UK

Contact author: christian.fullgrabe@ihr.mrc.ac.uk

Abstract

With the advent of cognitive hearing science, increased attention has been given to individual differences in cognitive functioning and their explanatory power in accounting for inter-listener variability in understanding speech in noise (SiN). The psychological construct that has received most interest is working memory (WM), representing the ability to simultaneously store and process information. Common lore and theoretical models assume that WM-based processes subtend speech processing in adverse perceptual conditions, such as those associated with hearing loss or background noise. Empirical evidence confirms the association between WM capacity (WMC) and SiN identification in older hearing-impaired listeners. To assess whether WMC also plays a role when listeners without hearing loss process speech in acoustically adverse conditions, we surveyed published and unpublished studies in which the Reading-Span test (a widely used measure of WMC) was administered in conjunction with a measure of SiN identification. The survey revealed little or no evidence for an association between WMC and SiN performance. We also analysed new data from 132 normal-hearing participants sampled from across the adult lifespan (18 to 91 years), for a relationship between Reading-Span scores and identification of matrix sentences in noise. Performance on both tasks declined with age, and correlated weakly even after controlling for the effects of age and audibility ($r = 0.39$, $p \leq 0.001$, one-tailed). However, separate analyses for different age groups revealed that the correlation was only significant for middle-aged and older groups but not for the young (< 40 years) participants.

Keywords: speech in noise, working memory, aging, normal hearing

1 Introduction

Recent years have seen an increased interest in the role of individual differences in cognitive functioning in speech and language processing and their interaction with different types of listening tasks and conditions. The psychological construct that has received the most attention in the emerging field of cognitive hearing science is working memory (WM), possibly because it has been shown to be involved in a wide range of complex cognitive behaviours (*e.g.* reading comprehension, reasoning, complex learning). WM can be conceptualised as the cognitive

system that is responsible for active maintenance of information in the face of ongoing processing and/or distraction. Its capacity (WMC) is generally assessed by so-called complex span tasks, requiring the temporary storage and simultaneous processing of information. For example, in one of the most widely used WM tasks, the Reading-Span task (Baddeley et al., 1985), visually presented sentences have to be read and their semantic correctness judged (processing component), while trying to remember parts of their content for recall after a variable number of sentences (storage component).

A growing body of evidence from studies using mainly older hearing-impaired (HI) listeners indeed confirms that higher WMC is related to better unaided and aided speech-in-noise (SiN) identification, with correlation coefficients frequently exceeding 0.50 (Lunner, 2003; Foo et al., 2007; Lunner & Sundewall-Thorén, 2007; Arehart et al., 2013). In addition, high-WMC listeners were less affected by signal distortion introduced by hearing-aid processing (*e.g.* frequency or dynamic-range compression).

Consistent with these results, models of speech/language processing have started incorporating active cognitive processes (Rönnberg et al., 2013; Heald & Nusbaum, 2014). For example, according to the Ease of Language Understanding model (Rönnberg et al., 2013), any mismatch between the perceptual speech input and the phonological representations stored in long-term memory disrupts automatic lexical retrieval, resulting in the use of explicit, effortful processing mechanisms based on WM. Both internal distortions (*i.e.*, related to the integrity of the auditory, linguistic, and cognitive systems) and external distortions (*e.g.* background noise) are purportedly susceptible to contribute to the mismatch. Consequently, it is assumed that WMC also plays a role when individuals with normal hearing (NH) have to process spoken language in acoustically adverse conditions.

However, Füllgrabe *et al.* (2015) recently failed to observe a link between Reading-Span scores and SiN identification in older listeners (≥ 60 years) with audiometrically NH (≤ 20 dB HL between 0.125 and 6 kHz), using a range of target speech (consonants and sentences), maskers (unmodulated and modulated noise, interfering babble), and signal-to-noise ratios (SNRs).

2 Literature survey

To assess the claim that individual variability in WMC accounts for differences in SiN identification even in the absence of peripheral hearing loss, we surveyed the literature for studies administering the Reading-Span test and a measure of SiN identification to participants with audiometrically NH. To ensure consistency with experimental conditions in investigations of HI participants, only studies presenting sentence material "traditionally" used in hearing research (*i.e.*, ASL, Hagerman, HINT, IEEE, QuickSIN, or Versfeld sentences) against co-located background maskers were considered. In addition, we only examined studies in which the effect of age was controlled for (either by statistically partialling it out or by restricting the analysis to a "narrow" age range), in order to avoid inflated estimates of the correlation between WMC and SiN tasks caused by the tendency for performance in both kinds of tasks to worsen with age. Figure 1 summarizes the results of this survey.

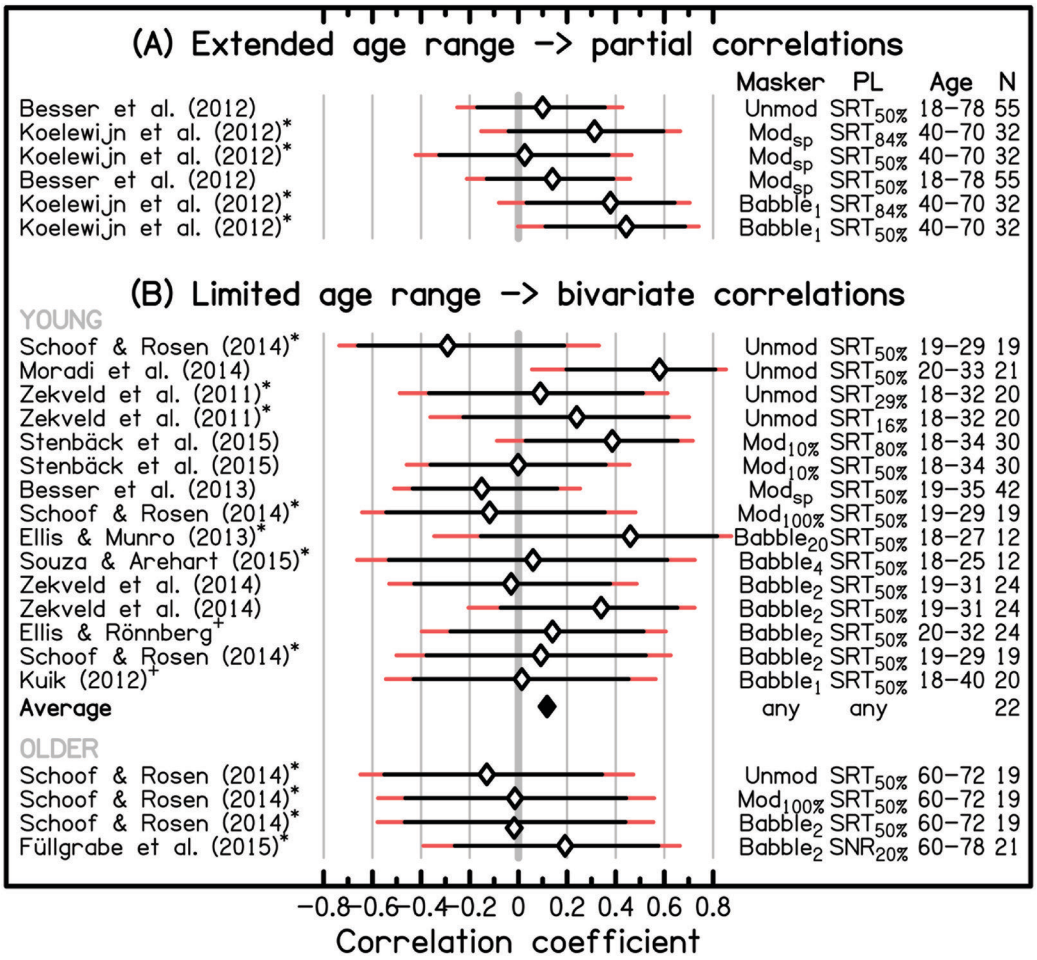


Figure 1. Comparison of Pearson correlation coefficients (diamonds) and associated 95% (black) and 99% (red) confidence intervals for studies investigating the association between WMC and speech-in-"noise" identification in NH participants after controlling for the effect of age by (A) computing partial correlations, or (B) using a limited age range. When necessary, the sign of the correlation was changed so that a positive correlation represents good performance on the two tasks. A weighted average for correlations based only on young NH listeners is provided (multiple *r* values for the same study sample are entered as their average). Source references (* indicates re-analysed published data; + indicates unpublished data, personal communication) and experimental (type of masker, Masker; performance level, PL) and participant (age range, Age; number of participants, N) details are given in the figure. Masker: Unmod – unmodulated noise, $Mod_{X\%\ or\ sp}$ – noise modulated by an X% sinusoidal amplitude modulation or a speech envelope, $Babble_X$ – X-talker babble. PL: $SRT_{X\%}$ – adaptive procedure tracking the speech reception threshold corresponding to X%-correct identification, $SNR_{X\%}$ – fixed SNR levels yielding, on average, X%-correct identification.

Correlation coefficients in the surveyed studies are broadly distributed, spanning almost half of the possible range of *r* values (*i.e.*, from –0.29 to 0.58). Confidence intervals (CIs) are generally large and include the null hypothesis in 21/25 and 24/25 cases for CIs of 95 and 99%, respectively, suggesting that these studies are not appropriately powered. For the relatively small number of studies included in this survey, there is no consistent trend for stronger correlations in more complex and/or informationally masking backgrounds or at lower SNRs, presumably corresponding to more adverse listening conditions.

Across studies restricting their sample to young (18-40 years) participants, the weighted average *r* value is 0.12, less than 2% of the variance in SiN identification. According to a power calculation, it would require 543 participants to have an 80% chance of detecting such a small effect with $p = 0.05$ (one-tailed)!

3 Analysis of Cohort Data for Audiometrically Normal-Hearing Participants

Given the mixed results from previous studies based on relatively small sample sizes, we re-analysed data from a subset of a large cohort of NH listeners taking part in another study.

3.1 Method

Participants were 132 native-English-speaking adults, sampled continuously from across the adult lifespan (range = 18-91 years). Older (≥ 60 years) participants were screened using the Mini Mental State Examination to confirm the absence of cognitive impairment. All participants had individual audiometric hearing thresholds of ≤ 20 dB HL at octave frequencies between 0.125 and 4 kHz, as well as at 3 kHz, in the test ear. Despite clinically "normal" audibility, the pure-tone average (PTA) for the tested frequency range declined as a function of age (r = 0.65, p ≤ 0.001, one-tailed). Since changes in sensitivity even in the normal audiometric range can affect SiN identification (Dubno & Ahlstrom, 1997), PTA is treated as a possible confounding variable in analyses involving the entire age group.

WMC was assessed by means of the computerized version of the Reading Span test (Rönnberg et al., 1989). Individual sentences were presented in a word-by-word fashion on a computer screen to be read aloud and judged as plausible or implausible. After three to six sentences, either the first or last word of each of the sentences had to be recalled. WMC corresponded to the number of correctly recalled words in any order.

SiN identification was assessed using the English version of the Matrix test (Vlaming et al., 2011). Each target sentence, presented monaurally at 70 dB SPL, followed a fixed syntactic structure (proper noun-verb-numeral-adjective-noun) but had a low semantic redundancy. The noise maskers had the same long-term spectrum as the target sentences and were either unmodulated or 100% sinusoidally amplitude modulated at 8 or 80 Hz. Target and masker were mixed together at SNRs ranging from −3 to −15 dB, and the mixture was lowpass-filtered at 4 kHz.

3.2 Results and Discussion

Identification scores were transformed into rationalized arcsine units (RAUs) and averaged across masker types and SNRs to reduce the effect of errors of measurement and to yield a composite intelligibility score representative of a range of test conditions.

Confirming previous results for audiometrically NH listeners (Füllgrabe et al., 2015), Reading-Span and SIN identification scores showed a significant decline with age, with Pearson's r = −0.59 and −0.68 (both $p \leq 0.001$, one-tailed), respectively. The scatterplot in Figure 2 shows that, considering all ages, performances on the tasks were significantly related to each other (r = 0.64, $p \leq 0.001$, one-tailed). This association remained significant after partialling out the effects of age and PTA (r = 0.39, $p \leq 0.001$, one-tailed), contrasting with the results of Besser *et al*. (2012), using a cohort including only a few (N=8) older (≥ 60 years) participants, but being roughly consistent with those reported by Koelewijn *et al*. (2012) for a cohort comprised of middle-aged and older (≥ 40 years) participants (see Fig. 1A).

To further investigate the age dependency of the association between WMC and SiN identification, participants were divided into four age groups: "Young" (range = 18-39 years, mean = 28 years; N = 32), "Middle-Aged" (range = 40-59 years, mean = 49 years; N = 26), "Young-Old" (range = 60-69 years, mean = 65 years; N = 40) and "Old-Old" (range = 70-91 years, mean = 77 years; N = 34). Separate correlational analyses for each age group revealed that the strength of the association differed across groups (see Fig. 2). Consistent with the overall trend seen in Figure 1, the correlation was weak and non-significant in the group of young participants (r = 0.18, p = 0.162, one-tailed). In contrast, the correlations were moderately strong and

significant in the three older groups (all $r \geq 0.44$, all $p \leq 0.011$, one-tailed). Comparing the different correlation coefficients, after applying Fisher's r-to-z transformation, revealed a significant difference between the Young and Old-Old group ($z = -1.75$, $p = 0.040$, one-tailed). There was no evidence for a difference in variance between these groups (Levene's test, $F_{(1,64)} < 1$, $p = 0.365$).

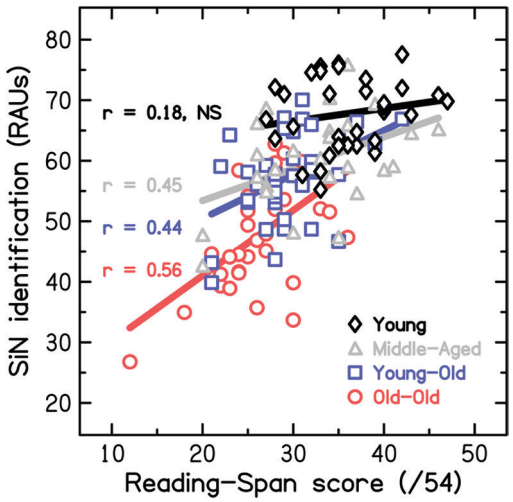


Figure 2. Scatterplot relating SiN identification averaged across background noises and SNRs to Reading-Span scores for the four age groups. The best linear fit to the data (thick lines) and associated bivariate Pearson correlation coefficients for each age group are given in the figure.

The age-related modulation of the strength of the correlation between WMC and SiN perception could be due to the different performance levels at which the age groups operated in this study (mean identification was 68, 60, 57, and 48 RAUs for the Young, Middle-Aged, Young-Old, and Old-Old group, respectively). However, when performance only for the two lowest SNRs (corresponding to 46 RAUs) was considered, WMC was still not associated with SiN identification in the young participants ($r = 0.04$, $p = 0.405$, one-tailed).

4 Conclusions

Taken together, the reported results fail to provide evidence that, in *acoustically* adverse listening situations, WMC (as measured by the Reading-Span test) is a reliable and strong predictor of SiN *intelligibility* in young listeners with normal hearing. The new data presented here suggest that WMC becomes more important with age, especially in the oldest participants. One possible explanation for this increasing cognitive involvement with age could be the accumulation of age-related deficits in liminary but also supraliminary auditory processing (e.g. sensitivity to temporal-fine-structure and temporal-envelope cues; Füllgrabe, 2013; Füllgrabe et al., 2015), resulting in under-defined and degraded internal representations of the speech signal, calling for WM-based compensatory mechanisms to aid identification and comprehension.

Our findings do not detract from the practical importance of cognitive assessments in the prediction of SiN identification performance in older HI listeners and the possible interaction between cognitive abilities and hearing-aid processing. Nor do they argue against the involvement of cognition in speech and language processing in young NH listeners *per se*. First, individual differences in WMC have been shown to explain some of the variability in performance in more linguistically complex task (such as in the comprehension of dynamic conversations; Keidser et al., 2015), presumably requiring memory or attentional/inhibitory processes associated with WMC (Conway et al., 2001; Kjellberg et al., 2008). Second, different cognitive measures, probing the hypothesized sub-processes of WM (e.g. inhibition, shifting, updating) or other domain-general cognitive primitives (*e.g.* processing speed) might prove to be better predictors of SiN processing abilities than the Reading-Span test.

In conclusion, and consistent with recent efforts to establish if and under which conditions cognitive abilities influence the processing of spoken language (e.g. Fedorenko, 2014; Heinrich & Knight, this volume), the current results caution against the assumption that WM necessarily supports SiN identification independently of the age and hearing status of the listener.

Acknowledgments

We would like to thank our colleagues who shared and reanalysed their data, and Dr. Oliver Zobay for his statistical advice. The MRC Institute of Hearing Research is supported by the Medical Research Council (grant number U135097130). This work was also supported by the Oticon Foundation (Denmark). CF is indebted to Prof. Brian Moore for granting access to the test equipment of his laboratory.

References

Arehart, K. H., Souza, P., Baca, R., & Kates, J. M. (2013). Working memory, age, and hearing loss: susceptibility to hearing aid distortion. *Ear and Hearing, 34*(3), 251-260.

Baddeley, A., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24*(1), 119-131.

Besser, J., Zekveld, A. A., Kramer, S. E., Rönnberg, J., & Festen, J. M. (2012). New measures of masked text recognition in relation to speech-in-noise perception and their associations with age and cognitive abilities. *Journal of Speech, Language, and Hearing Research, 55*(1), 194-209.

Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2013). How linguistic closure and verbal working memory relate to speech recognition in noise--a review. *Trends in Amplification, 17*(2), 75-93.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review, 8*(2), 331-335. doi: Doi 10.3758/Bf03196169

Dubno, J. R., & Ahlstrom, J. B. (1997). Additivity of multiple maskers of speech. In W. Jesteadt (Ed.), *Modeling Sensorineural Hearing Loss* (pp. 253-272). Hillsdale NJ: Lawrence Erlbaum Associates.

Ellis, R. J., & Munro, K. J. (2013). Does cognitive function predict frequency compressed speech recognition in listeners with normal hearing and normal cognition? *International Journal of Audiology, 52*(1), 14-22.

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology, 5*, 335.

Foo, C., Rudner, M., Rönnberg, J., & Lunner, T. (2007). Recognition of speech in noise with new hearing instrument compression release settings requires explicit cognitive storage and processing capacity. *Journal of the American Academy of Audiology, 18*(7), 618-631.

Füllgrabe, C. (2013). Age-dependent changes in temporal-fine-structure processing in the absence of peripheral hearing loss. *American Journal of Audiology, 22*(2), 313-315.

Füllgrabe, C., Moore, B. C., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience, 6*, 347.

Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 35.

Keidser, G., Best, V., Freeston, K., & Boyce, A. (2015). Cognitive spare capacity: evaluation data and its association with comprehension of dynamic conversations. *Frontiers in Psychology, 6*, 597.

Kjellberg, A., Ljung, R., & Hallman, D. (2008). Recall of words heard in noise. *Applied Cognitive Psychology, 22*(8), 1088-1098. doi: Doi 10.1002/Acp.1422

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology, 2012*, 865731.

Kuik, A. M. (2012). *Speech reception in noise: On auditory and cognitive aspects, gender differences and normative data for the normal-hearing population under the age of 40.* Bachelor's thesis, Vrije Universiteit Amsterdam, Amsterdam.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology, 42 Suppl 1*, S49-58.

Lunner, T., & Sundewall-Thorén, E. (2007). Interactions between cognition, compression, and listening conditions: effects on speech-in-noise performance in a two-channel hearing aid. *Journal of the American Academy of Audiology, 18*(7), 604-617.

Moradi, S., Lidestam, B., Saremi, A., & Rönnberg, J. (2014). Gated auditory speech perception: effects of listening conditions and cognitive capacity. *Frontiers in Psychology, 5*, 531.

Rönnberg, J., Arlinger, S., Lyxell, B., & Kinnefors, C. (1989). Visual evoked potentials: relation to adult speechreading and cognitive function. *Journal of Speech and Hearing Research, 32*(4), 725-735.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlstrom, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7*, 31.

Souza, P. E., & Arehart, K. H. (2015). Robust relationship between reading span and speech recognition in noise. *International Journal of Audiology*, In press.

Stenbäck, V., Hällgren, M., Lyxell, B., & Larsby, B. (2015). The Swedish Hayling task, and its relation to working memory, verbal ability, and speech-recognition-in-noise. *Scandinavian Journal of Psychology, 56*(3), 264-272.

Vlaming, M. S. M. G., Kollmeier, B., Dreschler, W. A., Martin, R., Wouters, J., Grover, B., Mohammadh, Y., & Houtgast, T. (2011). HearCom: Hearing in the Communication Society. *Acta Acustica United with Acustica, 97*(2), 175-192.

Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience, 8*, 88.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Festen, J. M., van Beek, J. H., & Rönnberg, J. (2011). The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise. *Ear and Hearing, 32*(6), e16-25.

Chapter 5

The Contribution of Auditory and Cognitive Factors to Intelligibility of Words and Sentences in Noise

Antje Heinrich and Sarah Knight
MRC Institute of Hearing Research, Nottingham, UK

Contact author: AH, antje.heinrich@ihr.mrc.ac.uk

Abstract

Understanding the causes for speech-in-noise (SiN) perception difficulties is complex, and is made even more difficult by the fact that listening situations can vary widely in target and background sounds. While there is general agreement that both auditory and cognitive factors are important, their exact relationship to SiN perception across various listening situations remains unclear. This study manipulated the characteristics of the listening situation in two ways: first, target stimuli were either isolated words, or words heard in the context of low- (LP) and high-predictability (HP) sentences; second, the background sound, speech-modulated noise, was presented at two signal-to-noise ratios. Speech intelligibility was measured for 30 older listeners (aged 62–84) with age-normal hearing and related to individual differences in cognition (working memory, inhibition and linguistic skills) and hearing ($PTA_{0.25\text{-}8kHz}$ and temporal processing). The results showed that while the effect of hearing thresholds on intelligibility was rather uniform, the influence of cognitive abilities was more specific to a certain listening situation. By revealing a complex picture of relationships between intelligibility and cognition, these results may help us understand some of the inconsistencies in the literature as regards cognitive contributions to speech perception.

Keywords: Speech-in-noise perception, listening situations, cognitive and auditory variables

1 Introduction

Speech-in-noise (SiN) perception is something that many listener groups, including older adults, find difficult. Previous research has shown that hearing sensitivity cannot account for all speech perception difficulties, particularly in noise (Schneider & Pichora-Fuller, 2000; Wingfield & Tun, 2007). Consequently, cognition has emerged as another key factor. While there is general agreement that a relationship between cognition and speech perception exists, its nature and extent remain unclear. No single cognitive component has emerged as being important for all listening contexts, although working memory, as tested by reading span, often appears to be important.

Working memory (WM) has no universally-accepted definition. One characterisation posits that WM refers to the ability to simultaneously store and process task-relevant information (Daneman & Carpenter, 1980). WM tasks may emphasize either storage or processing. Storage-heavy tasks, such as Digit Span and Letter-Number Sequencing (Wechsler, 1997), require

participants to repeat back material either unchanged or slightly changed. Processing-heavy tasks, such as the Reading Span task (Daneman & Carpenter, 1980), require a response that differs considerably from the original material and is only achieved by substantial mental manipulation.

The correlation between WM and speech perception, particularly in noise, tends to be larger when the WM task is complex i.e. processing-heavy (Akeroyd, 2008). However, this is only a general trend: not all studies show the expected correlation (Koelewijn et al., 2012), and some studies show significant correlations between WM and SiN perception even though the WM measure was storage-, not processing-heavy (Humes et al., 2006; Rudner et al., 2008). Why these inconsistencies occurred remains to be understood.

WM can also be conceptualised in terms of inhibition of irrelevant information (Engle & Kane, 2003), which has again been linked to speech perception. For instance, poor inhibition appears to increase susceptibility to background noise during SiN tasks (Janse, 2012). Finally, general linguistic competence – specifically vocabulary size and reading comprehension – has also been shown to aid SiN perception in some situations (Avivi-Reich et al., 2014).

Some of the inconsistencies in the relationship between SiN perception and cognition are most likely caused by varying combinations of speech perception and cognitive tasks. Like cognitive tasks, speech perception tasks can be conceptualised in different ways. Speech tasks may vary along several dimensions including the complexity of the target (e.g. phonemes vs. sentences), of the background signal (e.g. silence vs. steady-state noise vs. babble), and/or the overall difficulty (e.g. low vs. high signal-to-noise ratio) of the listening situation. It is difficult to know if and to what extent variations along these dimensions affect the relationship to cognition, and whether these variations may explain, at least in part, inconsistencies between studies. For instance, could differences in target speech help explain why some studies (Desjardins & Doherty, 2013; Moradi et al., 2014), using a complex sentence perception test, found a significant correlation between reading span and intelligibility, while another study (Kempe et al., 2012), using syllables and the same cognitive test, did not? The goal of this study is to systematically vary the complexity of the listening situation and to investigate how this variation affects the relationship between intelligibility and assorted cognitive measures.

Finally, it is important to note that a focus on cognitive contributions does not imply that auditory contributions to SiN perception are unimportant. Besides overall hearing sensitivity we also obtained a suprathreshold measure of temporal processing by measuring the sensitivity to change in interaural correlation. This task is assumed to estimate loss of neural synchrony in the auditory system (Wang et al., 2011), which in turn has been suggested to affect SiN perception.

2 Methods

2.1 Listeners

Listeners were 30 adults aged over 60 (mean: 70.2 years, SD: 6.7, range: 62-84) with age-normal hearing. Exclusion criteria were hearing aid use and non-native English language status.

Initial assessment

Average pure-tone air-conduction thresholds (left ear only) are shown in Figure 1A. The pure-tone average (PTA) of all measured frequencies was used as an individual measure of hearing sensitivity. In order to determine the presentation level for all auditory stimuli including cognitive tasks, speech reception thresholds were obtained using 30 sentences from the Adaptive Sentence List (MacLeod & Summerfield, 1990). The speech level was adaptively varied starting at 60 dB SPL and the average presentation level of the last two reversals of a three-

down, one-up paradigm with a 2dB step size was used. Presenting all stimuli at 30dB above that level was expected to account for differences in intelligibility in quiet.

2.2 Tasks

Speech Task

Sentences: Stimuli were 112 sentences from a recently developed sentence pairs test (Heinrich et al., 2014). This test, based on the SPIN-R test (Bilger et al., 1984), comprises sentence pairs with identical sentence-final monosyllabic words, which are more or less predictable from the preceding context (e.g. 'We'll never get there at this rate' versus 'He's always had it at this rate'). High and low predictability (HP/LP) sentence pairs were matched for duration, stress pattern, and semantic complexity, and were spoken by a male Standard British English speaker.

Words: A total of 200 words comprising the 112 final words from the sentence task and an additional 88 monosyllables were re-recorded using a different male Standard British English speaker.

Noise: All speech stimuli were presented in speech-modulated noise (SMN) derived from the input spectrum of the sentences themselves. Words were presented at signal-to-noise ratios (SNRs) of +1 dB and -2 dB, sentences at -4 dB and -7 dB. SNR levels were chosen to vary the overall difficulty of the task between 20% and 80% accuracy. Intelligibility for each of six conditions (words and LP/HP sentences at low and high SNRs) was measured.

Auditory task

Temporal processing task (TPT)

Duration thresholds were obtained for detecting a change in interaural correlation (from 0 to 1) in the initial portion of a 1-second broadband (0-10 kHz) noise presented simultaneously to both ears. A three-down, one-up 2AFC procedure with 12 reversals, of which the last 8 were used to estimate the threshold, was used. Estimates were based on the geometric mean of all collected thresholds (minimum 3, maximum 5).

Cognitive tasks

Letter-Number Sequencing (LNS)

The LNS (Wechsler, 1997) measures mainly the storage component of WM although some manipulation is required. Participants heard a combination of numbers and letters and were asked to recall the numbers in ascending order, then the letters in alphabetical order. The number of items per trial increased by one every 3 trials; the task stopped when all 3 trials of a given length were repeated incorrectly. The outcome measure was the number of correct trials.

Reading Span Task (RST)

The RST places greater emphasis on manipulation (Daneman & Carpenter, 1980). In each trial participants read aloud unconnected complex sentences of variable length and recalled the final word of each sentence at the end of a trial. The number of sentences per trial increased by one every 5 trials, from 2 to 5 sentences. All participants started with trials of 2 sentences and completed all 25 trials. The outcome measure was the overall number of correctly recalled words.

Visual Stroop

In a variant on the original Stroop colour/word interference task (Stroop, 1935) participants were presented with grids of 6 rows of 8 coloured blocks. In two grids (control), each of the 48 blocks contained "XXXX" printed in 20-pt font at the centre of the block; in another two grids

(experimental), the blocks contained a mismatched colour word (e.g. "RED" on a green background). In both cases, participants were asked to name the colour of each background block as quickly and accurately as possible. Interference was calculated by subtracting the time taken to name the colours on the control grids from the time taken to name the colours in the mismatched experimental grids.

Mill Hill Vocabulary Scale (MH)

The Mill Hill (Raven et al., 1982) measures acquired verbal knowledge in a 20-word multiple-choice test. For each word participants selected the correct synonym from a list of 6 alternatives. A summary score of all correct answers was used.

Nelson-Denny Reading Test (ND)

The Nelson-Denny (Brown et al., 1981) is a reading comprehension test containing eight short passages. Participants were given 20 minutes to read the passages and answer 36 multiple-choice questions. The outcome measure was the number of correctly answered questions.

2.3 Procedure

Testing was carried out in a sound-attenuated chamber using Sennheiser HD280 headphones. With the exception of the TPT all testing was in the left ear only. Testing took place over the course of two sessions around a week apart. TPT and cognitive tasks were split between sessions without a strict order. The word and sentence tasks were always tested in different sessions, with the order of word and sentence tasks counterbalanced across participants. Each sentence-final word was only heard once, either in the context of an HP or an LP sentence, and half the sentences of each type were heard with high or low SNR. Across all listeners, each sentence-final word was heard an equal number of times in all four conditions (sentence type x SNR). After hearing each sentence or word participants repeated as much as they could. Testing was self-paced. The testing set-up for the cognitive tasks was similar but adapted to task requirements.

3 Results

Fig 1B presents mean intelligibility for stimulus type and SNR. A 3 stimulus type (words, LP sentences, HP sentences) by 2 SNR (high, low) repeated-measures ANOVA showed main effects of type ($F(2, 50) = 192.55$ $p < 0.001$, LP < words < HP) and SNR ($F(1, 25) = 103.43$, $p < 0.001$) but no interaction ($F(2, 50) = 1.78$, $p = 0.18$), and suggested that a 3-dB decrease in SNR reduced intelligibility for all three stimuli types by a similar amount (12%). There was also a significant difference in intelligibility between LP and HP sentences.

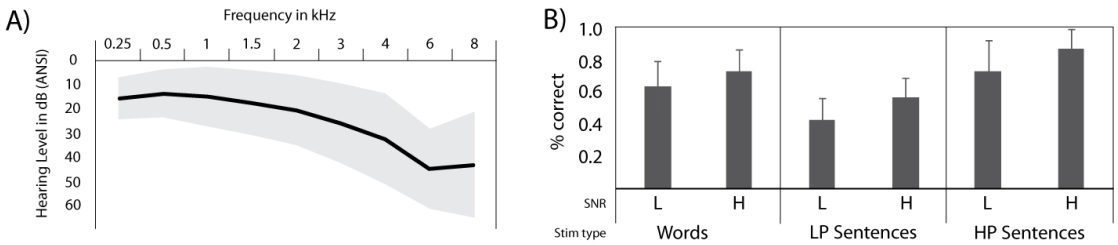


Figure 1. A) Audiometric thresholds (mean ± 1 SD) as a function of frequency for the left ear. B) Intelligibility (mean + 1SD) for three stimuli types (words, LP/HP sentences) presented at high (H) and low (L) SNR.

Table 1. (A) Significant p-values in linear mixed models estimating the effect of each auditory and cognitive variable on listening situations varying in type of target speech (words, LP/HP sentences) and SNR. WM = working memory, PTA = pure-tone average 0.25-8 kHz (left ear), LNS = Letter-Number Sequencing, RST = Reading Span Task, ND = Nelson-Denny reading comprehension. ME = main effect. **(B)** Pearson product-moment correlations between each auditory / cognitive variable and intelligibility in each of six listening situations. Words = monosyllables, LP = low-predictability, HP = high-predictability. Significant correlations are in italics.

		A. p-values			B. Pearson product-moment correlations					
					Words		LP sentences		HP sentences	
	var	ME	SNR *var	Type *SNR *var	Low SNR	High SNR	Low SNR	High SNR	Low SNR	High SNR
	PTA	<0.001	0.007	0.05	*-.57*	*-.60*	*-.66*	*-.53*	*-.69*	*-.61*
Auditory	TPT			0.04	-.18	*-.38*	-.33	-.20	-.29	-.26
Storage	LNS	0.003			.34	*.43*	*.54*	*.55*	.38	*.55*
Processing	RST		0.05		.16	.20	.33	.03	.36	.27
Inhibition	Stroop				.01	.06	-.08	.16	.10	.12
Vocab	MH				-.03	-.06	.07	.03	-.02	.22
Compr	ND				.06	.10	.30	.02	.17	.26

The effect of auditory and cognitive factors on intelligibility was examined in a series of separate linear mixed models (LMM with SSTYPE1). Each model included one auditory or cognitive variable, both as main effect and in its interactions with stimulus type (words, LP, HP) and SNR (L, H). The previously confirmed main effects for stimulus type and SNR were modelled but are not separately reported. Participants were included with random intercepts. Table 1A displays the p-values for all significant effects. Non-significant results are not reported. Table 1B displays bivariate correlations between each variable and the scores in each listening situation to aid interpretation. Note however that bivariate correlations do not need to be significant by themselves to drive a significant effect in an LMM.

The main effect of PTA reflected the fact that average audiometric threshold was negatively correlated with intelligibility in all listening situations. The interaction with SNR occurred because the correlations tended to be greater for low SNRs than high SNRs. This was particularly true for the sentence stimuli leading to the three-way interaction (Type * SNR * PTA). The Type*SNR*TPT interaction occurred because TPT thresholds showed a significant negative correlation with intelligibility for word stimuli at high SNR. The main effect of LNS occurred because a better storage-based WM score was beneficial for intelligibility in all tested listening situations. Examining the bivariate correlations to understand the interaction between WM processing-heavy RST and SNR suggests that a good score in the RST was most beneficial for listening situations with low SNR, although this relationship did not reach significance in any case. Intelligibility was not affected by inhibition, vocabulary size or reading comprehension abilities

4 Discussion

In this study we assessed individual differences in auditory and cognitive abilities and investigated their predictive value for SiN perception across a range of tasks. By systematically varying the characteristics of the listening situation we hoped to resolve inconsistencies in the cognitive speech literature regarding correlations between cognitive and speech tasks. By assessing multiple relevant abilities we also aimed to understand if and how the contribution of a particular ability varied across listening situation.

The results suggest that the importance of a particular variable often, but not always, depends on the listening situation. Hearing sensitivity (PTA) and a basic WM task correlated with

intelligibility in all tested situations. The results for PTA are somewhat surprising given that all speech testing was done at sensitivity-adjusted levels, which might have been expected to equate for PTA differences. The PTA measure may therefore capture some aspect of hearing that is not well represented by SRT.

The importance of WM for intelligibility is less surprising. However, the rather basic, storage-based WM task appeared to capture a more general benefit, at least in the tested listening situations, than the often-used RST. While the RST did predict intelligibility, the effect was stronger for acoustically difficult listening situations (low SNR). The data provide support for the notion that the relationship between RST and speech perception depends on the listening situations and that inconsistent results in the literature may have occurred because not all speech tasks engage WM processes enough to lead to a reliable correlation. Lastly, we showed an effect of individual differences in temporal processing on intelligibility, but this effect was limited to easily perceptible (high SNR) single words. Possibly, temporal information is most useful when the stimulus is clearly audible and no other semantic information is available.

These results suggest that different variables modulate listening in different situations, and that listeners may vary not only in their overall level of performance but also in how well they perceive speech in a particular situation depending on which auditory/cognitive abilities underpin listening in that situation and how successful the listener is at employing them.

Acknowledgments

This research was supported by BBSRC grant BB/K021508/1.

References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*.

Avivi-Reich, M., Daneman, M., & Schneider, B. A. (2014). How age and linguistic competence alter the interplay of perceptual and cognitive factors when listening to conversations in a noisy environment. *Frontiers Systems Neuroscience*. doi: 10.3389/fnsys.2014.00021

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research, 27*, 32-48.

Brown, J. I., Bennett, J. M., & Hanna, G. (1981). *The Nelson-Denny Reading Test*. Chicago, IL: Riverside.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25*, 975-979. doi: 10.1121/1.1907229

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behaviour, 19*, 450-466.

Desjardins, J. L., & Doherty, K. A. (2013). Age-Related Changes in Listening Effort for Various Types of Masker Noises. *Ear and Hearing, 34*(3), 261-272.

Engle, R. W., & Kane, M. J. (2003). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 44, pp. 145-199). San Diego, CA: Elsevier.

Heinrich, A., Knight, S., Young, M., Moorhouse, R., & Barry, J. (2014). *Assessing the effects of semantic context and background noise for speech perception with a new British English sentences set test (BESST)*. Paper presented at the British Society of Audiology Annual Conference, Keele University, UK.

Humes, L. E., Lee, J. H., & Coughlin, M. P. (2006). Auditory measures of selective and divided attention in young and older adults using single-talker competition. *Journal of the Acoustical Society of America, 120*(5), 2926-2937. doi: 10.1121/1.2354070

Janse, E. (2012). A non-auditory measure of interference predicts distraction by competing speech in older adults. *Aging, Neuropsychology and Cognition, 19*, 741-758. doi: 10.1080/13825585.2011.652590

Kempe, V., Thoresen, J. C., Kirk, N. W., Schaeffler, F., & Brooks, P. J. (2012). Individual Differences in the Discrimination of Novel Speech Sounds: Effects of Sex, Temporal Processing, Musical and Cognitive Abilities. *Plos One, 7*(11).

Koelewijn, T., Zekveld, A. A., Festen, J. M., Ronnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International journal of otolaryngology*, 1-11. doi: 10.1155/2012/865731

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24*, 29-43.

Moradi, S., Lidestam, B., Saremi, A., & Ronnberg, J. (2014). Gated auditory speech perception: effects of listening conditions and cognitive capacity. *Frontiers in Psychology*.

Raven, J. C., Raven, J., & Court, J. H. (1982). *Mill Hill Vocabulary Scale*. Oxford: Oxford Psychologists Press.

Rudner, M., Foo, C., Sundewall-Thorén, E., Lunner, T., & Rönnberg, J. (2008). Phonological mismatch and explicit cognitive processing in a sample of 102 hearing-aid users. *International Journal of Audiology, 47*(2), 91-98. doi: 10.1080/14992020802304393

Schneider, B. A., & Pichora-Fuller, M. K. (2000). Implications of perceptual deterioration for cognitive aging research. In F. I. M. Craik & T. A. Salthouse (Eds.), *The Handbook of Aging and Cognition* (pp. 155-219). Mahwah, N.J.: Erlbaum.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Wang, M., Wu, X., Li, L., & Schneider, B. A. (2011). The effects of age and interaural delay on detecting a change in interaural correlation: The role of temporal jitter. *Hearing Research, 275*, 139-149.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale-3rd Edition (WAIS-3®)*. San Antonio, TX: Harcourt Assessment.

Wingfield, A., & Tun, P. A. (2007). Cognitive supports and cognitive constraints on comprehension of spoken language. *Journal of the American Academy of Audiology, 18*, 548-558.

Chapter 6

Do Hearing Aids Improve Affect Perception?

Juliane Schmidt
Centre for Language Studies, Radboud University Nijmegen, Netherlands

Diana Herzog
Phonak AG, Staefa, Switzerland

Odette Scharenborg, Esther Janse
Centre for Language Studies, Radboud University Nijmegen, Netherlands
Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, Netherlands

Contact author: J.S., j.schmidt@let.ru.nl

Abstract

Normal-hearing listeners use acoustic cues in speech in order to interpret the speaker's emotional state. This study investigates how hearing loss affects the perception of the emotion dimensions arousal (aroused vs. calm) and valence (positive/negative attitude) in older adults using hearing aids. Affect ratings by 23 hearing aid users are compared for aided and unaided listening and are also compared to ratings by an age-matched group of 24 participants with age-normal hearing. More specifically, we investigate whether wearing a hearing aid improves the correlation between affect ratings and affect-related acoustic parameters.

The rating results for the evaluation of arousal show that the hearing aid users rated utterances as generally more aroused in the aided compared to the unaided condition. Both listening condition and hearing loss severity differences among hearing aid users changed the use of certain acoustic parameters. Compared to the reference group, hearing aid users showed minor differences in the use of intensity for arousal rating.

For valence, hearing loss severity did not influence ratings, and neither did listening condition (aided vs. unaided). For both emotion dimensions, ratings of hearing aid users in the aided condition did not generally differ from those of the participants with age-normal hearing.

Hearing impairment and the use of hearing aids thus matter particularly for the interpretation of arousal. Therefore, future studies on affect perception in hearing aid users should treat perception of arousal and valence separately.

Keywords: emotion perception, arousal, valence, acoustic parameters

1 Introduction

People use several sources of information to perceive and interpret emotions. Visual information, such as facial expressions, is most informative, but auditory cues in affective speech also provide cues for emotion perception. Auditory information can signal semantics (*what* is said) and prosody (*how* something is said). Prosodic cues are acoustic parameters in speech such as pitch, intensity, and tempo from which a normal-hearing listener may derive emotions from the speech signal (Banse & Scherer, 1996; Scherer, 2003; Coutinho & Dibben, 2013). Note that prosodic cues may alter the meaning of a spoken message, which becomes apparent when irony is expressed. For example, the meaning of an utterance like “I like roses” can either be

interpreted as positive (I do like roses) or negative (I do not like roses), depending on the applied prosody.

In an ideal communicative setting both visual and auditory information is available. Everyday communication settings, however, may frequently deprive the listener of visual information, e.g., when the listener is not able to face the speaker. Consequently, auditory cues to the emotional state of the speaker become especially important in these situations.

Moderate and severe hearing loss is treated with hearing aids in order to improve speech intelligibility, i.e., hearing aids are designed to improve the understanding of *what* is said. It is unclear, however, to what extent hearing aids sufficiently restore information needed for emotion perception in speech, i.e., whether hearing aids improve understanding of *how* something is said. Several studies with severely hearing-impaired children and adolescents indicate that aided hearing-impaired listeners perform poorly compared to normal-hearing listeners when rating emotion in speech (Most et al., 1993; Most & Aviner, 2009; Most & Michaelis, 2012). Importantly, affect perception differences between hearing-aid users and normal-hearing listeners were independent of hearing loss severity in the hearing-impaired participants. These findings on severe hearing loss obtained with children and adolescents, however, cannot be directly transferred to older adults using hearing aids, as it is well documented that there are age-related differences in the perception of affective prosody between younger and older participants (Paulmann et al., 2008; Dupuis & Pichora-Fuller, 2010).

To our knowledge, so far only the effect of mild hearing loss has been investigated in older adults (not wearing hearing aids). For this group, inconsistent findings have been reported concerning the link between individual hearing loss and the perception of affect. Orbelo and colleagues (Orbelo et al., 2005) report no effect of hearing sensitivity on affect perception, while Rigo and Lieberman (Rigo & Lieberman, 1989) found that low-frequency hearing loss ($PTA_{low(0.25,\ 0.5,\ 1\ kHz)}$ > 25 dB HL) impacted affect perception. Note that both studies used acted speech, where the lack of a global effect of hearing sensitivity on affect perception could be due to the more prototypical prosodic expression of affect in acted compared to natural speech (Scherer, 1986; Wilting et al., 2006). More extreme expressions of affect may be relatively easy to perceive, even for people with hearing loss (Grant, 1987) thus obscuring a possible influence of hearing sensitivity on affect perception.

The purpose of the current study is to investigate to what extent hearing aids improve affect perception, and how hearing loss in older adults influences affect perception. Participants will be tested on natural conversational speech stimuli in order to mimic realistic listening condition. To that end, older (bilateral) hearing aid wearers are tested with (aided condition) and without their hearing aid (unaided condition). Second, their aided performance is compared to a group of age-matched normal-hearing listeners. In particular, this study focuses on whether the degree of hearing loss and the use of hearing aids change listeners' sensitivity to the acoustic parameters cueing affect. We will therefore evaluate whether the use of hearing aids improves the correlation between these acoustic parameters and affect ratings of speech.

2 Methods

2.1 Participants

Two groups of older adults aged between 65 and 82 were tested. All participants were Swiss German native speakers and were compensated for their participation. A group of 23 older hearing aid users with bilaterally symmetric sensorineural hearing loss (M_{Age} = 73.5 years, SD_{Age} = 4.5; 17 men, 6 women) was recruited via the Phonak AG participant database. Participants have worn hearing aids bilaterally for at least two years. Furthermore, a group of 24 normal-hearing adults (M_{Age} = 70.3 years, SD_{Age} = 5.2; 10 men, 14 women) was recruited via the Phonak human resource department and a local senior club in Staefa, Switzerland.

Participant's hearing ability was tested by means of pure-tone audiometry (air conduction thresholds). The mean unaided pure-tone average (PTA) across 0.5, 1, 2, and 4 kHz for the hearing-impaired group was 49.8 dB HL (SD = 8.7, range: 32.5 – 68.8). The normal-hearing participants were expected to have age-normal thresholds, i.e., given their age, a downward sloping hearing loss in the frequencies above 1 kHz was expected, applying the statistical distribution of hearing thresholds as a function of age (ISO 7029:2000). Thresholds below the ISO's maximal pure-tone average threshold (across 0.5, 1, 2, and 4 kHz) at the age of 70 for men (PTA = 33.5 dB HL) and women (PTA = 26.0 dB HL) was considered as normal hearing. Additionally, participants underwent a brief cognitive screening test in order to exclude participants with mild cognitive impairment. We used the German version of the Montreal Cognitive Assessment Test (MOCA, Nasreddine et al., 2005), using a cutoff criterion of 67% (cf. Waldron-Perrine & Axelrod, 2012). The test was adjusted for hearing-impaired participants (Dupuis et al., 2015) by leaving out the tasks in which auditorily presented items had to be memorized (still applying the 67% cutoff criterion). None of the participants had to be excluded on the basis of this (modified) cognitive screening instrument.

2.2 Task and Procedure

Affect perception was tested in this study with the dimensional approach, in which participants do not label perceived emotions (as sounding "angry", or "sad"), but rather indicate the level of a specific emotion dimension on a rating scale. Participants rated the emotion dimensions arousal (calm vs. aroused) and valence (positive vs. negative attitude) separately. Stimuli were short utterances (audio-only) from an authentic and affectively-colored German conversational speech corpus (Grimm et al., 2008). Note that emotion inferences from speech correlate across languages; they particularly correlate for similar languages (cf. Scherer et al., 2001). Given the close relationship between German and Swiss German, the way affect is encoded in Swiss German is not expected to differ considerably from that in German as spoken in Germany. The corpus comes with mean reference values for the degree of arousal and valence for each utterance. These reference values had been collected with a simple 5-step pictorial rating tool (Bradley & Lang, 1994), ranging from -1 (calm/negative) to +1 (aroused/positive). The same rating tool was used to collect affective ratings in the current study. Utterances were selected to be neutral with respect to the content of what was said. Twenty-four utterances were selected for the arousal task (reference value range: -0.66 – 0.94) and 18 utterances were selected for the valence task (reference value range: -0.80 – 0.77). From these stimuli sets two randomized lists were created differing in the order in which the stimuli were presented for each emotion dimension.

Affect ratings were related to four acoustic parameters which are traditionally related to affective prosody: mean F0 (e.g., Hammerschmidt & Jürgens, 2007), mean intensity (e.g., Aubergé & Cathiard, 2003), global temporal aspects (Mozziconacci & Hermes, 2000), and spectral measures (e.g., Tamarit et al., 2008), the latter being related to vocal effort. For the current study, mean F0 and mean intensity were calculated for each utterance by averaging over the utterance using Praat (Boersma & Weenink, 2013). As a measure of tempo, articulation rate was calculated by dividing the number of syllables in the canonical transcription of the utterance by the file length, excluding pauses longer than 100 ms. Spectral slope is reflected in the spectral information described by the Hammarberg Index (Hammarberg et al., 1980). The Hammarberg Index is defined as the intensity difference between the maximum intensity in a lower frequency band [0 – 2000 Hz] versus that in a higher frequency band [2000 – 5000 Hz]. In this study, the Hammarberg Index was used as an energy distribution measure of relative intensity, averaged across the entire utterance.

Participants were comfortably seated in a sound-treated room and were tested in the free field. The pictorial rating tool was displayed on a computer screen and stimuli were presented via a single loudspeaker which was placed at head level in front of the participant (0° azimuth) with a distance of 1m from the participant's ears. Participants received written and oral instructions

and performed four practice trials before proceeding to the test stimuli of either rating task. Both rating tasks were completed at the participant's own pace. Utterances were rated one at a time and there was the possibility to replay the sentence.

All participants performed the rating tasks in two conditions. For the hearing aid users, these two conditions were with (aided) and without their hearing aids (unaided). The normal-hearing participants completed the tasks in a normal listening condition and in a condition with simulated hearing loss (data of the latter condition are not reported here). Order of rating tasks and listening conditions were counterbalanced over participants. The two different lists were used to present listeners with a different order of the stimuli in the two listening conditions. There was a short break between the two listening conditions and between the two rating tasks.

3 Results

The data were analyzed using R statistical software (R Development Core Team, 2008). Two comparisons were carried out separately for each emotion dimension (i.e., aided vs. unaided listening, and hearing-aid wearers vs. the normal-hearing group). The aim of the first analysis was two-fold. It investigated 1) whether the severity of the hearing loss modulates affect ratings and 2) whether wearing a hearing aid changes listeners' affect ratings, and, in particular, whether wearing a hearing aid makes listeners more sensitive to subtle differences in acoustic parameters. To that end, the affect ratings (the dependent variable) were compared for the hearing-impaired listeners in the aided and unaided conditions using linear mixed-effects regression analyses with random intercepts for stimulus and participant. The initial model allowed for a three-way interaction between listening condition (aided, unaided), individual hearing loss, and each of the acoustic parameters (mean F0, mean intensity, articulation rate, Hammarberg Index). Interactions and predictors that did not improve model fit (according to the Akaike Information Criterion) were removed using a stepwise exclusion procedure. Interactions were removed before simple effects, and those with the highest non-significant p-values were excluded first.

The second analysis investigated whether the use of a hearing aid restores affect perception to the level of normal-hearing older adults. Method and model-stripping procedure were identical to the first analysis. The initial model allowed for two-way interactions between group (HA users aided, normal hearing) and each of the four acoustic parameters.

3.1 Aided versus unaided listening

For arousal, mean intensity was found to be a strong cue for arousal rating ($\beta = 0.35$, SE = 0.08, $p < 0.001$): higher mean intensity is associated with higher ratings of arousal in both the aided and unaided conditions. Moreover, compared to the unaided condition (mapped on the intercept), ratings were generally higher in the aided condition ($\beta = 0.07$, SE = 0.02, $p < 0.001$). Moreover, significant interactions between listening condition and articulation rate ($\beta = 0.04$, SE = 0.02, $p < 0.05$) and the measure of vocal effort ($\beta = 0.08$, SE = 0.02, $p < 0.001$) were observed: even though vocal effort and articulation rate did not influence ratings in the unaided condition, they did in the aided condition. In the unaided condition, those with poorer hearing had lower ratings ($\beta = -0.08$, SE = 0.03, $p < 0.05$) than those with better hearing, but this was less the case in the aided condition ($\beta = 0.06$, SE = 0.02, $p < 0.01$). Furthermore, those with poorer hearing associated increases in F0 ($\beta = 0.03$, SE =0.01, $p < 0.01$) and in articulation rate ($\beta = 0.02$, SE = 0.01, $p < 0.05$) more with higher arousal than those with better hearing across listening conditions.

For valence, a significant simple effect of mean F0 ($\beta = -0.31$, SE = 0.06, $p < 0.001$) was found, i.e., higher pitch is associated with lower ratings, i.e., higher pitch was associated with more negative ratings. None of the other acoustic parameters was predictive of the valence ratings. Importantly, no effects for listening condition and hearing loss were observed: valence ratings

are independent of whether the participants wore their hearing aids or not and independent of individual hearing loss.

3.2 Hearing-aid users (aided condition) versus normal-hearing controls

As in the previous arousal analysis, a significant simple effect for mean intensity (β = 0.39, SE = 0.07, p = 0.001) was found: higher mean intensity was associated with higher arousal ratings. Although ratings of the hearing-aid users did not generally differ from the normal-hearing participants (i.e., from the group mapped on the intercept, β = -0.02, SE = 0.05, p = 0.75), use of mean intensity differed between the two listener groups: hearing-aid users respond more to differences in intensity than participants with age-normal hearing (β = 0.04, SE = 0.02, $p < 0.05$).

For valence, as in the previous analysis, pitch was associated with lower valence ratings as shown by a simple effect of mean F0 (β = -0.30, SE = 0.08, $p < 0.01$). No other acoustic parameters were predictive of ratings of valence. There was no effect of group, nor any interactions between group and the acoustic parameters.

4 Discussion

This study aimed to investigate whether the use of a hearing aid restores affect perception to the level of older adults with age-normal hearing. Furthermore, our study investigated to what extent hearing aids modify sensitivity to the acoustic parameters cueing affect in older hearing aid users. Moreover, we investigated the role of hearing loss severity, particularly in unaided listening.

Hearing aid users wearing their hearing aid generally showed the same pattern of affect ratings as participants with age-normal hearing, particularly for the valence dimension. However, for arousal ratings, those wearing a hearing aid actually used intensity information more than participants in the reference group. This may be because hearing in the reference group was normal for their age, but still implied higher thresholds for the high frequencies. Consequently, by using a hearing aid, hearing aid users have become more sensitive, at least to some acoustic differences, than the reference group.

For the hearing-impaired, the use of the hearing aid changed the pattern of acoustic parameters serving as cues for arousal. Two acoustic parameters that did not generally cue arousal in the unaided condition did cue arousal in the aided condition, namely articulation rate and the Hammarberg Index (measuring vocal effort). Note also that, across the aided and unaided conditions, whether or not listeners used alternative cues to interpret arousal (i.e., other cues than intensity) was modulated by hearing loss severity: hearing-impaired listeners with more severe degrees of hearing loss made more use of articulation rate and F0. This suggests that changes in sensory input quality will make listeners rely on different cues for their interpretation of arousal. In other words, gradually acquired hearing loss will make listeners rely on different cues, but restoring their hearing by means of a hearing aid will also change which cues they rely on for their interpretation of arousal. Possibly, the reason for using articulation rate is different for the poorest hearing (across listening conditions) than for those who start to use it in the aided condition. In a related study by Schmidt, Scharenborg, and Janse (Schmidt et al., in preparation) in which older adults with mild hearing loss were tested who were not wearing hearing aids, only mean intensity emerged as a significant predictor of arousal. While older listeners with mild hearing loss rely on just one acoustic parameter when rating arousal, older listeners with severe hearing loss may rely on several acoustic parameters in order to compensate for the lost information. Note again that this does not hold for valence, where F0 is the only prosodic cue listeners are using.

In sum, the current study underlines the importance of hearing aids in the rehabilitation of affect perception in older hearing-impaired listeners. Wearing hearing aids particularly matters for the interpretation of rating prosodic information related to arousal. Due to this difference between

emotion dimensions, future studies on affect perception in hearing aid users should treat perception of arousal and valence separately.

Acknowledgments

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7-PEOPLE-2011-290000.

References

Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication*, *40*, 87–97.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer. www.praat.org/. Accessed 10 April 2013.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Coutinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. *Cognition & Emotion*, *27*(4), 658–684.

Dupuis, K., & Pichora-Fuller, M. K. (2010). Use of affective prosody by young and older adults. *Psychology and Aging*, *25*(1), 16–29.

Dupuis, K., Pichora-Fuller, M. K., Chasteen, A. L., Marchuk, V., Singh, G., & Smith, S. L. (2014). Effects of hearing and vision impairments on the Montreal Cognitive Assessment. *Aging, Neuropsychology, and Cognition: A Journal on Normal and Dysfunctional Development*, 22(4), 413-437.

Grant, K. W. (1987). Identification of intonation contours by normally hearing and profoundly hearing-impaired listeners performance. *Journal of the Acoustical Society of America*, *82*(4), 1172–1178.

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (pp. 865–868). Hannover, Germany.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica*, *90*, 441–451.

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice : Official Journal of the Voice Foundation*, *21*(5), 531–40.

International Organization for Standardization (2000). Acoustics— Statistical Distribution of Hearing Thresholds as a Function of Age. ISO 7029. Geneva: ISO.

Most, T., & Aviner, C. (2009). Auditory, visual, and auditory-visual perception of emotions by individuals with cochlear implants, hearing aids, and normal hearing. *Journal of Deaf Studies and Deaf Education*, *14*(4), 449–464.

Most, T., & Michaelis, H. (2012). Auditory, visual, and auditory–visual perceptions of emotions by young children with hearing loss versus children with normal hearing. *Journal of Speech, Language, and Hearing Research*, *55*(4), 1148–1162.

Most, T., Weisel, A., & Zaychik, A. (1993). Auditory, visual and auditory–visual identification of emotions by hearing and hearing-impaired adolescents. *British Journal of Audiology*, *27*, 247–253.

Mozziconacci, S. J. L., & Hermes, D. J. (2000). Expression of emotion and attitude through temporal speech variations. In *Sixth International Conference on Spoken Language Processing* (pp. 373–378). Beijing, China.

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., Chertkow, H. (2005). http://www.mocatest.org/. Accessed 28 November 2013.

Orbelo, D. M., Grim, M. A., Talbott, R. E., & Ross, E. D. (2005). Impaired comprehension of affective prosody in elderly subjects is not predicted by age-related hearing loss or age-related cognitive decline. *Journal of Geriatric Psychiatry and Neurology*, *18*(1), 25–32.

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language*, *104*(3), 262–269.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. Accessed 4 March 2013.

Rigo, T. G., & Lieberman, D. A. (1989). Nonverbal sensitivity of normal-hearing and hearing-impaired older adults. *Ear and Hearing*, *10*(3), 184–189.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227–256.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92.

Tamarit, L., Goudbeek, M., & Scherer, K. (2008). Spectral slope measurements in emotionally expressive speech. In *Proceedings for ISCA ITRW Speech Analysis and Processing for Knowledge Discovery*. Aalborg, Denmark.

Waldron-Perrine, B., & Axelrod, B. N. (2012). Determining an appropriate cutting score for indication of impairment on the Montreal Cognitive Assessment. *International Journal of Geriatric Psychiatry*, *27*(11), 1189–1194.

Wilting, J., Krahmer, E., & Swerts, M. (2006). Real vs. acted emotional speech. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech)* (pp. 805–808). Pittsburgh, Pennsylvania.

Chapter 7

Suitability of the Binaural Interaction Component for Interaural Electrode Pairing of Bilateral Cochlear Implants

Hongmei Hu, Birger Kollmeier, and Mathias Dietz

Medizinische Physik, Cluster of Excellence "Hearing4all", Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Contact author: H. Hu, hongmei.hu@uni-oldenburg.de

Abstract

Although bilateral cochlear implants (BiCIs) have succeeded in improving the spatial hearing performance of bilateral CI users, the overall performance is still not comparable with normal hearing listeners. Limited success can be partially caused by an interaural mismatch of the place-of-stimulation in each cochlea. Pairing matched interaural CI electrodes and stimulating them with the same frequency band is expected to facilitate binaural functions such as binaural fusion, localization, or spatial release from masking. It has been shown in animal experiments that the magnitude of the binaural interaction component (BIC) derived from the wave-eV decreases for increasing interaural place of stimulation mismatch. This motivated the investigation of the suitability of an electroencephalography-based objective electrode-frequency fitting procedure based on the BIC for BiCI users. A 61 channel monaural and binaural electrically evoked auditory brainstem response (eABR) recording was performed in 7 MED-EL BiCI subjects so far. These BiCI subjects were directly stimulated at 60% dynamic range with 19.9 pulses per second via a research platform provided by the University of Innsbruck (RIB II). The BIC was derived for several interaural electrode pairs by subtracting the response from binaural stimulation from their summed monaural responses. The BIC based pairing results are compared with two psychoacoustic pairing methods: interaural pulse time difference sensitivity and interaural pitch matching. The results for all three methods analyzed as a function of probe electrode allow for determining a matched pair in more than half of the subjects, with a typical accuracy of ±1 electrode. This includes evidence for statistically significant tuning of the BIC as a function of probe electrode in human subjects. However, results across the three conditions were sometimes not consistent. These discrepancies will be discussed in the light of pitch plasticity versus less plastic brainstem processing.

Keywords: binaural interaction component, electrically evoked auditory brainstem response, bilateral cochlear implant, pitch matching, interaural pulse time difference, interaural electrode pairing

1 Introduction

Bilateral cochlear implantation seeks to restore the advantages of binaural hearing to the profound deaf by providing binaural cues that are important for binaural perception, such as binaural fusion, sound localization and better detection of signals in noise. Most bilateral cochlear implant (BiCI) users have shown improvements compared to their ability when only

one CI was used. However, compared to normal hearing (NH) individuals, the average performance of BiCI users is still worse and has a large variability in performance amongst them (Majdak et al., 2011; Litovsky et al., 2012; Goupell et al., 2013; Kan et al., 2013). One likely reason for the worse performance of BiCI users is the interaural electrodes mismatch between two CIs because of different surgery insertion depth or different implant length. Since the inputs to the NH binaural system from the two ears can be assumed to be well matched and binaural brainstem neurons are comparing only by interaurally place matched inputs, it is very important to determine interaural electrode pairs for frequency matching the electrode arrays in the two ears to compensate for any differences in the implanted cochleae. This interaural electrode pairing (IEP) is expected to become even more relevant in the future, to better exploit recent time information preserving coding strategies and with the advent of truly binaural cochlear implants, which will preserve, enhance, and/or optimize interaural cues.

It has been suggested that physiological measures of binaural interactions (e.g., evoked potentials) will likely be required to accomplish best-matched interaural electrode pairs (Pelizzone et al., 1990). More recently, the binaural interaction component (BIC) of the electrically evoked auditory brainstem responses (eABRs) have been obtained in both animals (Smith & Delgutte, 2007) and humans (He et al., 2010; Gordon et al., 2012). Smith & Delgutte (2007) proposed a potential way by using evoked potentials to match interaural electrode pairs for bilaterally implanted cats. Their study shows that the interaural electrode pairings that produced the best aligned IC activation patterns were also those that yielded the maximum BIC amplitude. More recently, He et al. (2010) observed some evidence of a BIC/electrode-offset interaction at low current levels. In another follow up study, they used the same electroencephalography (EEG) procedure to examine whether the BIC amplitude evoked from different electrode pairings correlated with an interaural pitch comparison task (He et al., 2012). Their results show that there is no significant correlation between results of BIC measures and interaural pitch comparisons on either the individual or group levels. Gordon et al. (2012) demonstrated that binaural processing in the brainstem of children using bilateral CI occurs regardless of bilateral or unilateral deafness; it is disrupted by a large but not by a small mismatch in place of stimulation. All these studies suggest that BIC could be a potentially approaches for electrode pairing, especially for pediatrics, however, it is not accurate enough in all existing studies of human subjects.

In order to tackle the issue of accuracy, a 61-channel monaural and binaural eABR recording setup was developed together with a multi-step offline post-processing strategy specifically for eABR signals. Further, to address the question of method validity the BIC based pairing results are compared with two psychoacoustic pairing methods: interaural pulse time difference sensitivity and interaural pairwise pitch comparison.

2 Methods

Seven BiCIs (three male and four female; mean age = 53 year) participated in this study. All of them were post-lingual at their onset of bilateral severe-to-profound sensorineural hearing loss and were using MED-EL implant systems. The voluntary informed written consent was obtained with the approval of the Ethics Committee of the University of Oldenburg. Table 1 shows some information and the tested electrodes of the 7 participants. The intra-cochlear electrodes of the MED-EL CI are numbered from 1 to 12 in an apical to basal direction. Here the electrode on the right or the left implant is named Rx or Lx. For example, L4 refers to the 4th electrode on the left implant.

A research platform was developed for psychoacoustic testing and eABR recording (Hu et al., 2014). The electrical stimuli were presented via a research interface box (RIB II, manufactured at University of Innsbruck, Austria) and a National Instruments I/O card. The monaural or

bilaterally synchronized electrical pulses were applied to the CI, bypassing the behind-the-ear unit.

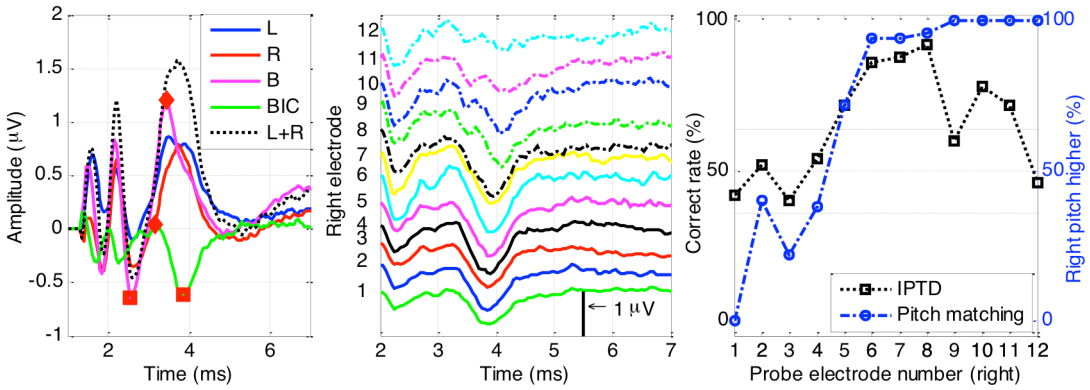
The experiment consisted of three psychophysical pretests, accompanied by three IEP methods. Pretest 1 was to determine the participant's maximum comfortable level (MCL) and the hearing threshold (HL) of the reference electrode (For more details about the stimuli, please refer to the last paragraph of Section 2). From these two values the dynamic range (DR) was obtained. In the second pretest the interaural loudness-balanced level for each electrode pair was determined. The presentation level of the reference CI electrode was fixed at 60% of DR, while the presentation level of the contralateral probe CI electrode was adapted according to the participant's response. The stimulation levels judged by the participant to be equally loud in both sides were saved and used in the pairwise pitch comparison and EEG recording procedures. In the interaural pairwise pitch comparison task interaural electrode pairs were stimulated sequentially. The participants were asked to indicate in which interval the higher pitch was perceived. This experiment was repeated 50 times per electrode pair. As a third pre-test, a lateralization task was performed on each electrode pair to ensure that a single, fused auditory image was perceived in the center of the head (Kan et al., 2013). The current level and electrode pairs that allowed for a fused and central image were again stored and employed for the second IEP method: IPTD discrimination. Two-interval trials, randomly located with left-leading ($\mathrm{IPTD} = -\mathrm{T}/2$) and right-leading stimuli ($\mathrm{IPTD} = \mathrm{T}/2$), were presented for the electrode pair using a constant stimulus procedure, for a fixed $\mathrm{IPTD} = \mathrm{T}$. The participant was required to indicate whether the stimulus in the second interval was perceived either to the left or the right of the first interval.

In the EEG measurement session, eABRs were differentially recorded from Ag/AgCl electrodes of a customized equidistant 63-channel braincap (Easycap) with an electrode at FPz serving as the ground, and the midline cephalic location (Cz) as the physical reference (Hu et al., 2015). The 63 scalp electrodes were connected to the SynAmps RT amplifier system (Neuroscan). Channel 49 (sub-Inion) and channel 59 (Inion) were primary interest in this study. Electrode impedances were kept below 10 KΩ. The sweeps were filtered by an analog antialiasing-lowpass with a corner frequency of 8 kHz, digitized with 20 kHz sampling rate via a 24 bit A/D convertor. The artifact rejection was turned off during the recording, since filtering, artifact analysis and averaging were done offline. 2100-3000 single sweeps for all electrode conditions were recorded in random order on a sweep-by-sweep basis. During the recording, the participants seated in a recliner and watched silent subtitled movies within an electrically shielded sound-attenuating booth.

The stimulus was a train of charge-balanced biphasic pulses presented at a rate of 19.9 pulses per second (pps) (He et al., 2010), with 50 or 60 μs phase duration, and 2.1 μs interphase gap presented repeatedly via monopolar stimulation mode. A phase duration of 60 μs was used only when the subject could not reach the MCL with 50 μs. In psychophysical sessions 1 to 4, a 10-pulse chain about 500 ms time duration was used. In the IPTD test, each interval was a 500 ms stimulus and there was 300 ms pause between the two intervals. For the EEG recoding, a continuous electrical pulse train was presented with a 5 ms trigger sent 25 ms before each the CI stimulation onset.

3 Results

The BIC is defined as the difference between the eABR with binaural stimulation (B) and the sum of the eABRs obtained with monaural stimulation (L+R): BIC = B - (L+R). For both eABR and BIC, the latency was defined as the time position of the respective peaks. The amplitude was defined as the difference between the positive peak and the following trough amplitude.



(a) (b) (c)

Figure 1. EABRs and the BICs of subject S5 and the reference electrode is L4: (a) EABRs of left CI stimulated only, right CI stimulated only, both CI stimulated simultaneously, L+R and the BIC. The probe electrode is R1. The y-axis is their amplitude values in μV. The electric artifact was lined out, wave eV and BIC are visible at approximately 3.6 and 4.1 ms, respectively. (b) The results of the BIC across 12 electrode pairs. (c) The IEP results based on the best IPTD performance and pitch comparison. The x-axis is the number of the probe electrodes in the right implant.

Figure 1 (a) shows the morphology of the eABR and the BIC of subject S5 with reference electrode L4 and probe electrode R1, where the electric artifact was already removed (Hu et al., 2015). In general, wave eI was not observed due to the stimulation pulse artifact from the implant. The amplitude of wave eV was larger than an acoustically evoked ABR amplitude, and latency which occurs approximately 3.6 ms after the onset of the stimulus is shorter because the electrical stimulus directly activates the neural pathway (Starr & Brackmann, 1979; Pelizzone et al., 1989). Figure 1 (b) shows the results of the BIC across 12 electrode pairs (R1-R12), for S5 with reference electrode of L4. The latency of the BIC is about 4 ms, and the BIC amplitude is increasing first and then decreasing after electrode 6 (R6). Figure 1 (c) shows the IEP results based on the best IPTD performance and the pitch comparison. The open circles on the dash-dotted line indicate the percentage of how often the probe electrode (right side) resulted in a higher pitch percept than the reference electrode (left side) in the pitch matching experiment. The squares on the dashed line are the correct rates of lateralization judgment in the IPTD experiment. The x-axis indicates the probe electrode number in the right implant. The matched electrodes are R4.4 (pitch comparison), R6-R8 (best IPTD performance), and R6 (largest BIC amplitude). In contrast to the common assumption that pitch-matched pairs maximize sensitivity to binaural cues, for this subject the pairs with best IPTD sensitivity are 1.6-3.6 electrodes away from (interpolated) electrode identical pitch perception. The BIC based method indicates electrode R6 to be matched with L4, which is closer to the IPTD based match for this subject. For S5, there are three pairs that have a high IPTD sensitivity (R6 -R8). They are not significantly different from each other, rendering the assignment of a unique matched pair difficult. This is consistent with the previous finds that ITD JNDs do not change much until there is a 3–6 mm mismatch (van Hoesel & Clark, 1997). Note that the pitch matched electrode pair does not even yield a significantly above chance IPTD sensitivity.

The results from the other methods and of the other subjects are not shown because of the space limitations. The general trends were: 1) Pairwise pitch comparison results highly depend on the participant's clinical frequency-place fitting map. The matched pair indicates a similar frequency range in the map, except for one subject which was implanted only 9 months before the experiment. 2) Both BIC and IPTD based pairing results indicate a large 2-3 electrode mismatch for some subjects. There is clear correlation between the mismatches indicated by these two

methods. 3) In line with Poon et al. (2009) there is no correlation between the pitch matched pairs and the IPTD matched pair and in line with He et al. (2012) there is no correlation between the pitch matched pairs and the BIC matched pair.

4 Discussion

With the increasing number of bilateral CI users and new product development towards binaural CIs, it becomes increasingly important to find ways of matching the electrode arrays in the two ears to compensate for the differences between the two implanted cochleae because of insertion depth, electrode lengths, or neural survival. Current clinical cochlear implant fitting strategies generally treat each implant monaurally and allocate a generic frequency range to each electrode based on the electrode number often without concern about the actual position and the mismatching between the two implants. In a very few positive exceptions interaural pitch comparison techniques are employed. This study investigated the suitability of three IEP methods which have previously been tested in research studies but mostly only one method in isolation. In general, clear monaural and binaural eABRs and BICs were obtained in most of the subjects. All but one subject showed nicely tuned pitch matching, IPTD sensitivity, and BIC as a function of electrode number. However, the pairing results were not always consistent across methods. The IPTD paring results of some subjects are spanned across several electrodes, which is consistent with the previous finds that ITD JNDs do not changed much until there is a 3–6 mm mismatch compared to the best matched place condition (van Hoesel & Clark, 1997).

Interaural pitch comparison shows some predictive value in selecting interaural electrode pairs, but it may not fully compensate for any underlying differences between the two implanted cochleae. In line with previous studies (van Hoesel & Clark, 1997; Long et al., 2003; Poon et al., 2009) the pitch matched pair does not guarantee best IPTD sensitivity. All the CI subjects in this study have had at least 9 months bilateral CI experience prior to the experiment so acclimatization to the electrode-frequency-map or brain plasticity may explain the good pitch tuning curve. There is no correlation between the BIC amplitudes and results of interaural pitch matching as well, which is consistent with (He et al., 2012).

For the interpretation of the BIC, the straightest assumption is that if the reference electrode and the probe electrode are cochlear place matched, they stimulate similarly “tuned” auditory-nerve fibers that eventually result in binaural interaction in the brain stem and thus in a larger BIC than the other pairs. However, in case of a different neural survival across the ears or within one ear, also non-place matched electrodes may result in the maximum BIC. Further studies with more reference electrodes are necessary to test this assumption. This may also partially explain the similar ITD sensitivity of the neighboring electrodes around the best matched electrode.

In summary, all three methods can obtain reasonable interaural electrode pairing results. There is substantial variability in all the measurements both within and between individual CI users. There are pros and cons for all the three methods and there is no gold standard to judge which one is better. The acclimatization to the frequency to place map of the patient’s CI seems unignorably in the pitch perception. Pitch matching data may be ‘misleadingly good’ because of acclimatization or brain plasticity which does not affect the IPTD and the BIC based matching. Knowledge about the location of the implant inside the cochlea maybe helpful in judging the pairing results. Longitudinal measurements are necessary in the future to investigate the plasticity of the binaural perception in the BiCI users. The BIC is a promising candidate for electrode pairing of BiCI subjects, especially for pediatric fitting.

Acknowledgments

This work was funded by European Union under the Advancing Binaural Cochlear Implant Technology (ABCIT) grant agreement (No. 304912). We thank our ABCIT project partners at the

UCL Ear Institute David McAlpine and Torsten Marquardt for inspirational discussions when planning this study. The authors are grateful to Stefan Strahl (MED-EL) for RIB II support and Stephan Ewert and Tom Campbell for their help during method development. We thank Karsten Plotz, Katrin Bomke (Klinik für Phoniatrie und Pädaudiologie, Ev. Krankenhaus Oldenburg), and Stefan Dazert (CI-Zentrum Ruhrgebiet, Universitätsklinik Bochum) for recruiting CI participants. The authors also thank all CI participants for joining the study and for giving us valuable feedback.

References

Gordon, K. A., Salloum, C., Toor, G. S., van Hoesel, R., & Papsin, B. C. (2012). Binaural interactions develop in the auditory brainstem of children who are deaf: effects of place and level of bilateral electrical stimulation. *The Journal of Neuroscience, 32*(12), 4212-4223.

Goupell, M. J., Stoelb, C., Kan, A., & Litovsky, R. Y. (2013). Effect of mismatched place-of-stimulation on the salience of binaural cues in conditions that simulate bilateral cochlear-implant listening. *Journal of the Acoustical Society of America, 133*(4), 2272-2287.

He, S., Brown, C. J., & Abbas, P. J. (2010). Effects of stimulation level and electrode pairing on the binaural interaction component of the electrically evoked auditory brain stem response. *Ear and Hearing, 31*(4), 457-470.

He, S., Brown, C. J., & Abbas, P. J. (2012). Preliminary results of the relationship between the binaural interaction component of the electrically evoked auditory brainstem response and interaural pitch comparisons in bilateral cochlear implant recipients. *Ear and Hearing, 33*(1), 57-68.

Hu, H., Kollmeier, B., & Dietz, M. (2015). Reduction of stimulation coherent artifacts in electrically evoked auditory brainstem responses. *Biomedical Signal Processing and Control, xx*(in Press), xx-xx.

Hu, H., Ewert, S., Campbell, T., Kollmeier, B., & Dietz, M. (2014). *An Interaural Electrode Pairing Clinical Research System for Bilateral Cochlear Implants*. Paper presented at the the 2nd IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP' 14), Xi'an, China.

Kan, A., Stoelb, C., Litovsky, R. Y., & Goupell, M. J. (2013). Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users. *Journal of the Acoustical Society of America, 134*(4), 2923-2936.

Litovsky, R. Y., Goupell, M. J., Godar, S., Grieco-Calub, T., Jones, G. L., Garadat, S. N., Agrawal, S., Kan, A., Todd, A., Hess, C., & Misurelli, S. (2012). Studies on bilateral cochlear implants at the University of Wisconsin's Binaural Hearing and Speech Laboratory. *Journal of the American Academy of Audiology, 23*(6), 476-494.

Long, C. J., Eddington, D. K., Colburn, H. S., & Rabinowitz, W. M. (2003). Binaural sensitivity as a function of interaural electrode position with a bilateral cochlear implant user. *Journal of the Acoustical Society of America, 114*(3), 1565-1574.

Majdak, P., Goupell, M. J., & Laback, B. (2011). Two-dimensional localization of virtual sound sources in cochlear-implant listeners. *Ear and Hearing, 32*(2), 198-208.

Pelizzone, M., Kasper, A., & Montandon, P. (1989). Electrically evoked responses in Cochlear implant patients. *Audiology, 28*(4), 230-238.

Pelizzone, M., Kasper, A., & Montandon, P. (1990). Binaural interaction in a cochlear implant patient. *Hearing Research, 48*(3), 287-290.

Poon, B. B., Eddington, D. K., Noel, V., & Colburn, H. S. (2009). Sensitivity to interaural time difference with bilateral cochlear implants: Development over time and effect of interaural electrode spacing. *The Journal of the Acoustical Society of America, 126*(2), 806-815.

Smith, Z. M., & Delgutte, B. (2007). Using evoked potentials to match interaural electrode pairs with bilateral cochlear implants. *Journal of the Association for Research in Otolaryngology, 8*(1), 134-151.

Starr, A., & Brackmann, D. E. (1979). Brain stem potentials evoked by electrical stimulation of the cochlea in human subjects. *The Annals of otology, rhinology, and laryngology, 88*(4 Pt 1), 550-556.

van Hoesel, R. J., & Clark, G. M. (1997). Psychophysical studies with two binaural cochlear implant subjects. *Journal of the Acoustical Society of America, 102*(1), 495-507.

Chapter 8

Binaural Loudness Constancy

John F. Culling, Helen Dare
School of Psychology, Cardiff University, Cardiff, UK

Contact author: JFC, CullingJ@cf.ac.uk

Abstract

In binaural loudness summation, diotic presentation of a sound usually produces greater loudness than monaural presentation. However, experiments using loudspeaker presentation with and without earplugs find that magnitude estimates of loudness are little altered by the earplug, suggesting a form of loudness constancy. We explored the significance of controlling stimulation of the second ear using meatal occlusion as opposed to the deactivation of one earphone. We measured the point of subjective loudness equality (PSLE) for monaural vs. binaural presentation using an adaptive technique for both speech and noise. These stimuli were presented in a reverberant room over a loudspeaker to the right of the listener, or over lightweight headphones. Using the headphones, stimuli were either presented dry, or matched to those of the loudspeaker by convolution with impulse responses measured from the loudspeaker to the listener position, using an acoustic manikin. The headphone response was also compensated. Using the loudspeaker, monaural presentation was achieved by instructing the listener to block the left ear with a finger. Near perfect binaural loudness constancy was observed using loudspeaker presentation, while there was a summation effect of 3-6 dB for both headphone conditions. However, only partial constancy was observed when meatal occlusion was simulated. These results suggest that there may be contributions to binaural loudness constancy from residual low frequencies at the occluded ear as well as a cognitive element, which is activated by the knowledge that one ear is occluded.

Keywords: Loudness, Binaural, Summation, Constancy

1 Introduction

When the same sound is presented to both ears, it is perceived to be louder than when it is presented to one ear only (Fletcher & Munson, 1933). As a result, the loudness of diotically and monaurally presented stimuli, usually tones or noise, has been found to be equivalent when the monaural stimulus is between 3 and 10 dB more intense (Fletcher & Munson, 1933; Reynolds & Stevens, 1960; Zwicker & Zwicker, 1991; Sivonen & Ellermeier, 2006; Whilby et al., 2006; Edmonds & Culling, 2009). Models of loudness therefore incorporate this phenomenon in various ways (Zwicker & Scharf, 1965; Moore & Glasberg, 1996; Moore and Glasberg, 2007). These models can be used in hearing-aid fitting (e.g. Moore & Glasberg, 2004), so hearing aids fitted using these models will have relatively reduced gain when fitted bilaterally in order to achieve a comfortable maximum loudness level. However, some recent studies have cast doubt on the ecological validity of the loudness summation literature.

Cox & Gray (2001) collected loudness ratings (7 categories from “very soft” to “uncomfortable”) for speech at different sound levels when listening monaurally and binaurally. These two modes of listening were compared by using either one or two earphones and, using a loudspeaker, by

occluding one ear with a plug and ear muff. They found that results with one or two earphones produced a conventional loudness summation effect, whereby the mean rating was substantially higher at each sound level for binaural presentation. However, when listening to an external source in the environment (the loudspeaker) there was much less summation effect: occlusion of one ear had little effect on the loudness ratings compared to listening binaurally. This experiment showed for the first time that the loudness of external sounds may display constancy across monaural and binaural listening modes. However, the methods used were clinically oriented and difficult to compare with conventional psychophysical measurements.

Epstein & Florentine (2009, 2012) conducted similar tests, but using standard loudness estimation procedures and speech (spondees) either with or without accompanying video of the speaker's face. They also observed loudness constancy, but only when using the audiovisual presentation. Their tentative conclusion was that loudness constancy may only occur using stimuli of relatively high ecological validity. Ecological validity may be enhanced when an external source is used, when that source is speech, particularly connected speech rather than isolated words, and when accompanied by coherent visual cues. Since all these conditions are fulfilled when listening to someone in real life, the phenomenon of loudness constancy can be compellingly, if informally, demonstrated by simply occluding one ear with a finger when listening to someone talk; most people report that no change in loudness is apparent. The present study was inspired by this simple demonstration technique.

A limitation of previous studies is that the use of ear plugs and muffs means that different listening modes cannot be directly compared. It is commonplace in psychophysical testing to match the loudness of different stimuli by listening to them in alternation, but it is impractical to insert/remove an ear plug between one presentation interval and another. As a result comparisons are made over quite long time intervals. In contrast, a finger can be applied to the meatus in less than a second, so it is possible to perform a 2-interval, forced-choice procedure provided that the monaural interval is always the same one so that the listener can learn the routine of blocking one ear for the same interval of each trial. The present experiment adopted this technique and also explored the idea that ecological validity plays a key role by using either speech or noise as a sound source and by creating a close physical match between stimuli presented from a loudspeaker and those presented using headphones.

2 Methods

2.1 Stimuli

There were six conditions, comprised of two stimulus types (speech/noise) and three presentation techniques. The speech stimuli were IEEE sentences (Rothauser et al., 1969) and the noise stimuli were unmodulated noises filtered to have the same long-term excitation pattern (Moore & Glasberg, 1987) as the speech. These two stimulus types were presented (1) dry over headphones, (2) from a loudspeaker to the listeners' right or (3) virtually through lightweight open-backed headphones (Sennheiser HD414). Monaural presentation was achieved in (1) and (3) by deactivating the left earphone for the second interval of each trial and in (2) by asking the listener to block their left ear with a finger for the second interval of each trial. For virtual presentation, the stimuli were convolved with binaural room impulse responses (BRIRs) recorded from a manikin (B&K HATS 4100) sitting in the listeners' position and wearing the lightweight headphones with the loudspeaker 1 m to the right in a standard office room with minimal furnishing (reverberation time, RT60 = 650 ms). Impulse responses were also recorded from the headphones and used to derive inverse filters to compensate for the headphone-to-microphone frequency response. Sound levels for the different stimuli were equalised at the right ear of the manikin and, for the reference stimuli, were equivalent to 57 dB(A) as measured at the head position with a hand-held sound-level meter.

2.2 Procedure

12 undergraduate-students with no known hearing impairments took part in a single 1-hour session. Initially, 5 practice trials were presented, for which the listeners were simply required to learn the routine of blocking their left ear during the one-second inter-stimulus interval between two bursts of speech-shaped noise presented from a loudspeaker. During the experiment, the ordering of six conditions was randomly selected for each participant.

For each condition, listeners completed two loudness discrimination threshold tasks. These adaptive thresholds served to bracket the point of subjective loudness equality (PSLE): one was the 71% threshold for identifying when a monaural stimulus was louder than a binaural one, and the other was the 71% threshold for identifying when the monaural stimulus was quieter than a binaural one. For a given condition, the two adaptive tracks started with the adapted (monaural) sound level 15 dB above and below the level of the reference (binaural) stimulus, respectively. In each task, the listeners were required to identify the louder of two stimuli in a 2-down, 1-up adaptive procedure with six reversals. The two thresholds were averaged to yield a PSLE for that condition.

Within each trial the first interval was always the reference stimulus, while the adapted stimulus was presented in the second interval. The adapted stimulus was thus always the monaural one.

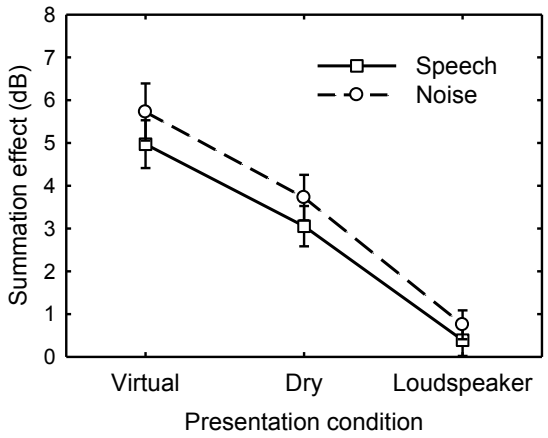


Figure 1. The mean summation effect in Expt. 1 in each of the three presentation conditions and for speech and for noise sources. Error bars are one standard error.

3 Results

Figure 1 shows that there was a strong effect of the presentation condition ($F(2,22)=22.1$, $p<0.001$). The summation effect is less than 1 dB in the Loudspeaker condition, in which monaural presentation was achieved by occluding the meatus with one finger, but is 3-6 dB in the Virtual and Dry conditions, in which monaural presentation was achieved by deactivating one earphone. Summation was also significantly smaller using speech than using noise as a source signal ($F(1,11) =5.6$, $p<0.05$).

4 Discussion

The lack of summation in the Loudspeaker condition supports the observations made by Cox & Gray (2001) and by Epstein & Florentine (2009, 2012), that, when monaural presentation is achieved by occluding one ear in a sound field, binaural summation can be dramatically reduced. The present results demonstrate this phenomenon for the first time using direct comparisons between successive stimuli. Consistent with suggestions that ecological validity is important to the effect, it was a little larger for speech than for noise, but this effect was rather small (0.8 dB).

In order to match the headphone and loudspeaker conditions physically, virtual presentation was used in a reverberant room. This approach resulted in a compelling impression of sound source externalisation for the headphone condition, which may have been accentuated by the presence of a loudspeaker at the appropriate position. A post-experiment test showed that listeners perceived the virtual binaural stimuli as emanating from the loudspeaker rather than the headphones. A consequence of this procedure was that when the left headphone was deactivated for monaural presentation, the externalisation of the sound collapsed, so that listeners found themselves comparing one sound that appeared to be from the loudspeaker with another that, although still reverberant and from the same side, appeared to come from the headphones. It is unclear whether this effect may have caused the apparent enhancement of the loudness summation effect in the Virtual condition.

In each of the studies in which binaural summation has been largely abolished, the second ear has been occluded by some means. The lack of summation could have two logical causes. First, there may be some residual sound energy at the occluded ear that is sufficient to maintain its contribution to binaural loudness. Second, there may be some cognitive element which generates loudness constancy through the knowledge that one ear is occluded. It is important, therefore to detemine the amount of residual sound and the role it might play. Cox & Gray used a combination of ear plug and muff and reported that this combination produced a threshold shift of 23 dB at 250 Hz and 52 dB at 4 kHz, indicating that the residual sound was substantially attenuated. Epstein & Florentine used only an ear plug and reported the resulting attenuation caused a threshold shift of 20-24 dB at 1 kHz. In the present study, a finger was used. In order to determine the likely degree of attenuation achieved, we recorded binaural room impulse responses BRIRs with and without an experimenter's finger over the meatus. This was not possible with the B&K HATS 4100 because it has the microphone at the meatus, so KEMAR was used. Several repeated measurements found a consistent attenuation of ~30 dB across most of the frequency spectrum, but progressively declining at low frequencies to negligible levels at 200 Hz and below. Assuming this measurement gives a reasonably accurate reflection of the change in cochlear excitation when a human listener blocks the meatus, the question arises whether this residual sound might be sufficient to support summation.

First, we conducted an analysis of the virtual stimuli by applying the Moore & Glasberg (1996) loudness model. Since similar effects were observed with noise as with speech, it was thought unnecessay to employ the model for time-varying sounds (Moore and Glasberg, 1997). Despite presentation of stimuli from one side, the difference in predicted loudness at each unoccluded ear differed by only 4%. The interaural level difference was small due to the room reverberation. When the left ear was occluded, the predicted loudness at this ear fell by 52%. Older models assume that binaural loudness is the sum of that for each ear, but more recent models predict less binarual summation by invoking an inhibitory mechanism (Moore and Glasberg, 2007). In either case, attenuation caused by meatal occlusion leading to a 52% drop at one ear seems more than adequate to observe a strong summation effect.

While modelling can be persuasive, we wanted to be sure that the relatively unattenuated low frequencies were not responsible for a disproportionate contribution to binaural loudness. A second experiment examined this possibility by simulating occlusion.

5 Methods

Eight listeners took part in a similar experiment, but using a new presentation condition that simulated the use of a finger to block the meatus. One of the BRIRs collected with a finger over the ear of a KEMAR was convolved with the source signals to form the monaural stimulus of a Virtual Finger condition. This condition was contrasted with a replication of the previous Virtual condition in which monaural presentation was achieved by silencing the left ear. This condition is now labelled Silence, so that the condition labels reflect the method of attenuating the left ear.

For consistency, all BRIRs for the second experiment were collected using KEMAR. The binaural stimulus for each condition was identical. The same speech and noise sources were used.

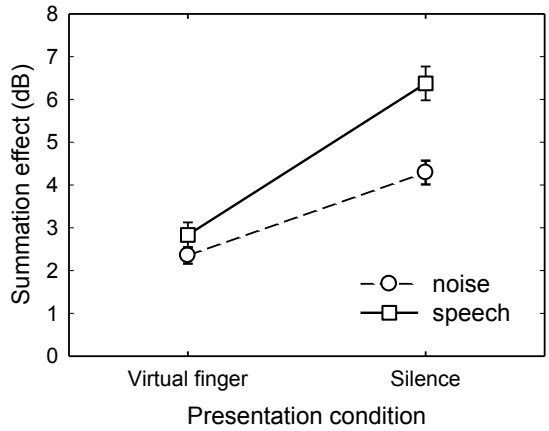


Figure 2. The mean summation effect in Experiment 2 in each of the presentation conditions and for speech and for noise sources. Error bars are one standard error.

6 Results

Figure 2 shows the mean summation effect from each of the four conditions of the second experiment. The Silence condition produces a comparable summation effect to the similarly constructed Virtual condition of Expt. 1. However, while the Virtual Finger condition does not result in an abolition of this summation effect, it does result in a substantial reduction in its size ($F(1,7)=102$, $p<0.001$). In contrast to Expt. 1, the noise stimulus produces a smaller summation effect than the speech stimulus ($F(1,7)=22$, $p<0.005$). The interaction fell short of significance.

7 Discussion

Since the Silence and Virtual-Finger conditions of experiment 2 produced significantly different summation effects, it seems that a portion of the loudness constancy effect observed in experiment 1 may have been mediated by the residual sound at the occluded ear. Both the Silence and the Virtual-Finger conditions produced a collapse in externalisation for the monaural case, so it seems unlikely that the difference in effect can be attributed to this cause.

On the other hand, 2.5-3 dB of summation was still observed in the Virtual-Finger condition, whereas less that 1 dB was observed in the Loudspeaker condition of experiment 1. It appears, therefore, that the listeners' awareness of meatal occlusion, which would only be present in the first case, may still play a role. A major caveat to these conclusions is that KEMAR was not designed to realistically simulate the effects of bone-conducted sound, which likely plays a major role in listening in a sound field with an occluded ear. While probe microphones in the ear canal might more accurately record the stimulation at the tympanic membrane, they would still not capture the stimulation occurring at the cochlea.

The small effect of stimulus type observed in experiment 1, was reversed in experiment 2. Taking these small and inconsistent effects together, we found no evidence that binaural loudness constancy is greater for more ecologically valid stimuli such as connected speech. Indeed, for both speech and noise, occlusion of the ear led to little reduction in loudness (a small summation effect), suggesting that loudness constancy occurs independently of stimulus type. It should be noted, however, that quite prominent reverberation was present in our experiment, which reliably cued the existence of an external sound source.

8 Conclusions

Experiment 1 demonstrated for the first time that binaural loudness constancy can be observed in a loudness matching task using direct comparisons of loudness between an occluded and unoccluded second ear. Experiment 2 showed that this effect could be partially mediated by the residual sound at the occluded ear, but the remaining effect would seem attributable to the listeners' awareness of the occlusion.

References

Cox, R. M., & Gray, G. A. (2001). Verifying loudness perception after hearing aid fitting. American Journal of Audiology, 10, 91–98.

Edmonds, B. A, & Culling, J. F. (2009). Interaural correlation and the binaural summation of loudness. Journal of the Acoustical Society of America, 125(6), 3865–70.

Epstein, M., & Florentine, M. (2009). Binaural loudness summation for speech and tones presented via earphones and loudspeakers. Ear and Hearing, 30(May), 234–237.

Epstein, M., & Florentine, M. (2012). Binaural loudness summation for speech and tones presented via earphones and loudspeakers. Journal of the Acoustical Society of America, 131, 3981–3988.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. Journal of the Acoustical Society of America, 5, 82–108.

Moore, B. C. J., & Glasberg, B. R. (1987). Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns, Hearing Research, 28, 209–225.

Moore, B. C. J., & Glasberg, B. R. (1996). A revision of Zwicker's loudness model. Acustica, 82, 335-345.

Moore, B. C. J., & Glasberg, B. R. (1996). A model for the prediction of thresholds, loudness, and partial loudness. Journal of the Audio Engineering Society, 45, 224-240.

Moore, B. C. J., & Glasberg, B. R. (2004). A revised model of loudness perception applied to cochlear hearing loss. Hearing Research, 188, 70–88.

Moore, B. C. J., & Glasberg, B. R. (2007). Modelling binaural loudness. Journal of the Acoustical Society of America, 121, 1604-1612.

Reynolds, G. S. & Stevens, S. S. (1960). Binaural summation of loudness. Journal of the Acoustical Society of America, 32. 1337-1344.

Sivonen, V. P., & Ellermeier, W. (2006). Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation. Journal of the Acoustical Society of America, 119, 2965–2980.

Whilby, S., Florentine, M., Wagner, E., & Marozeau, J. (2006). Monaural and binaural loudness of 5- and 200-ms tones in normal and impaired hearing. Journal of the Acoustical Society of America, 119, 3931–3939.

Zwicker, E., & Scharf, B. (1965). A model of loudness summation. Psychological Review, 72, 3–26.

Zwicker, E., & Zwicker, U. T. (1991). Dependence of binaural loudness summation on interaural level differences, spectral distribution and temporal distribution. Journal of the Acoustical Society of America , 89, 756–764.

Chapter 9

Intelligibility for Binaural Speech with Discarded Low-SNR Speech Components

Esther Schoenmaker, Steven van de Par

Acoustics Group, Cluster of Excellence "Hearing4All", Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

Contact author: ES, esther.schoenmaker@uni-oldenburg.de

Abstract

Speech intelligibility in multitalker settings improves when the target speaker is spatially separated from the interfering speakers. A factor that may contribute to this improvement is the improved detectability of target-speech components due to binaural interaction, in analogy to the Binaural Masking Level Difference (BMLD). This would allow listeners to hear target speech components within specific time-frequency intervals that have a negative SNR, similar to the improvement in the detectability of a tone in noise when these contain disparate interaural difference cues. To investigate whether these negative-SNR target-speech components indeed contribute to speech intelligibility a stimulus manipulation was performed where all target components were removed when local SNRs were smaller than a certain criterion value. It can be expected that for sufficiently high criterion values, target speech components will be removed that do contribute to speech intelligibility. For spatially separated speakers, assuming that a BMLD-like detection advantage contributes to intelligibility, degradation in intelligibility is expected already at criterion values below 0 dB SNR. However, for collocated speakers it is expected that higher criterion values can be applied without impairing speech intelligibility. Results show that degradation of intelligibility for separated speakers is only seen for criterion values of 0 dB and above, indicating a negligible contribution of BMLD-like processing in multitalker settings. These results show that the spatial benefit is related to a spatial separation of speech components at positive local SNRs rather than to a BMLD-like detection improvement for speech components at negative local SNRs.

Keywords: binaural listening, masking release, speech intelligibility

1 Introduction

When listening to speech in a noisy background, higher speech intelligibility is measured in the case of spatially separated speech and noise sources, as opposed to when all sources are collocated. The benefit of this spatial separation has been attributed to binaural processing that enables a better signal detection, as can be measured by binaural masking level differences (BMLDs). These BMLDs describe the difference in detection thresholds that can be measured when both the target tone and masking noise are presented interaurally in phase as compared to a reversed phase of the tone in one ear. In the latter case thresholds are significantly lower.

Similar interaural relations of the stimuli can be achieved by presenting the noise from a spatial position directly in front of the listener and the signal from a different spatial location. The

benefit achieved by the spatial separation relative to spatial collocation is referred to as spatial release from masking. Since spatial release from masking also had been measured for speech targets, a connection was hypothesized between the spatial benefit in speech intelligibility and the tone detection advantage that had been measured as BMLD.

Levitt & Rabiner (1967) used this hypothesis to successfully predict the spatial improvement in speech intelligibility. The underlying assumption of their model is that a BMLD effect acts within each frequency band, which leads to a within-band detection advantage. This was modeled as a reduction of the level of the noise floor by an amount that equals the size of the BMLD for that particular frequency band, resulting in an improved signal-to-noise ratio (SNR). In this manner previously undetectable speech elements become audible and can thus contribute to speech intelligibility. More recent models of speech intelligibility also make use of the concept of within-band improvement of SNR to model the effect of spatial release from masking (e.g., Beutelmann et al., 2010; Lavandier et al., 2012; Wan et al., 2014).

Whereas Levitt & Rabiner (1967) proposed their model for speech in a noise background, we are interested in modulated interference like speech. Since speech shows strong modulations both in time and frequency, the level of the interferer will not be constant over time like the stationary noise was in Levitt & Rabiner's model. Therefore we consider it more appropriate to assume a BMLD effect at the level of spectro-temporal regions rather than frequency channels. Thus we extend the original hypothesis and postulate that a BMLD effect might take place at the level of individual spectro-temporal units and that this detection advantage would be responsible for the spatially improved speech intelligibility.

The aim of this study is to investigate whether a BMLD-like effect, that would lead to improved detection of speech elements in a spatially separated source configuration, indeed is responsible for improved speech intelligibility in the presence of interfering speech.

We will test the contribution of a BMLD-like effect by measuring speech intelligibility while deleting spectro-temporal regions of the target signal for which the SNR does not meet a local criterion. Doing this we assume that, if certain speech elements contribute to speech intelligibility, the performance on a speech intelligibility task will be lower after deleting them. At low values of the SNR criterion this manipulation will lead to deletion of spectro-temporal regions at which either the target energy was masked by the presence of high masker energy or hardly any target energy was present at all. At such values not much impact on speech intelligibility is expected. Deletion of target speech components at increasing SNRs, however, will start to affect speech intelligibility at some point.

The question of interest is whether the effect of the removal of spectro-temporal regions will follow a different course for collocated and spatially separated source configurations. More specifically, it will be interesting to observe at which minimum value of the SNR criterion speech intelligibility will be affected. Should a BMLD-like effect indeed act at the level of small spectro-temporal regions, then one would expect an impairment of speech intelligibility to occur already at criterion SNRs of approximately –15 dB (i.e. diotic threshold – maximum BMLD). When target speech is deleted in spectro-temporal regions with local SNR values between –15 and 0 dB, it should be expected that speech intelligibility decreases for separated, but not for collocated sources, since those speech components would in any case be inaudible in a collocated configuration. There should thus exist a range of criterion values for which speech intelligibility will be affected by deletion of elements at lower SNRs in spatially separated conditions but not in spatially collocated conditions. This implies that over this same range of SNR criteria the spatial release from masking, as measured by the difference in performance between separated and collocated source configurations, should decrease.

2 Methods

2.1 Stimuli

German sentences taken from the Oldenburg Sentence Test (OLSA, Wagener et al., 1999) and spoken by a male speaker served as target speech. Each sentence consisted of five words with the syntactic structure *name – verb – numeral – adjective – object*. For each word type 10 alternatives were available, the random combination of which yields syntactically correct but semantically unpredictable sentences. Test lists consisted of 20 sentences each.

The interfering speech consisted of two streams of ongoing speech spoken by two different female talkers, and was obtained from two audio books in the German language. Any silent intervals exceeding 100 ms were cut from the signals to ensure a natural-sounding continuous speech stream without pauses (Giannakopoulos, 2010). All target and masker signals were available at a sampling rate of 44.1 kHz.

A total of 200 masker samples for each interfering talker were created by randomly cutting segments of 3.5-s duration from the preprocessed audio book signals. The onsets and offsets of the segments were shaped by 200-ms raised cosine ramps. Since the longest target sentences were about 2.8 s long, this ensured that the entire target sentence was masked by interfering speech. Masker fragments that did not contain German language or that could not be considered ongoing speech were discarded altogether.

The target sentence was padded with leading and trailing zeros to match the length of the interfering signals. Some time roving was applied to the onset of the target speech by randomly varying the number of leading zeros between 25% and 75% of the total number of zeros.

The relative levels of the three signals were set based on the mean RMS of the two stereo channels after spatialization. The two interfering signals were equalized to the same level, whereas the target speech (before application of the binary mask, see next section) was presented at a level of –8 dB relative to each single interferer. This resulted in a global SNR of approximately –11 dB. The sound level was set to 62 dB SPL for a single interferer. After signal manipulation as will be described below, the three signals were digitally added which resulted in a total sound level of approximately 65 dB SPL.

All speech signals were presented from virtual locations in the frontal horizontal plane. The spatial positions were rendered with the help of head-related transfer functions (HRTF), that had been recorded according to Brinkmann et al. (2013) using a Cortex MK2 head-and-torso simulator at a source-to-head distance of 1.7 m in an anechoic chamber.

Two different speaker configurations were used in this experiment, one with spatially separated speakers and one with collocated speakers. The target speech was presented from 0° azimuth, i.e. directly in front of the listener, in both configurations. In the collocated configuration the two sources of masking speech were presented from this same location. In the spatially separated configuration one interfering speaker was presented from the location 60° to the left and the other speaker from 60° to the right of the target.

2.2 Target Signal Manipulation

In order to investigate the contributions of target speech at various levels of local SNR, the target speech needed to be manipulated.

The first step in the stimulus manipulation consisted of the calculation of the local, i.e. spectro-temporal, SNR of the three-speaker stimulus. Spectrograms were calculated from 1024-point Fast Fourier Transforms on 1024-samples long 50%-overlapping square root Hann-windowed segments of the signals. This was performed for each of the two stereo channels of a spatialized speech signal to accommodate interaural level differences. This resulted in a pair of left and right spectrograms for each signal, from which one estimate of the spectro-temporal power

distribution needed to be calculated. This was achieved by computing the squared mean of the absolute values of each set of two (i.e. left and right) corresponding Fourier components and repeating this for the complete spectro-temporal plane (Figure 1, left column).

Subsequently, the resulting spectral power distributions were transformed to the equivalent rectangular bandwidth (ERB) scale. This was achieved by grouping power components that fell within the frequency range of one ERB. The time-frequency (T-F) units thus created covered a fixed time length of 23 ms (i.e. 1024 samples) and a frequency band of 1 ERB width. Within such a T-F unit the power estimates of spectrogram components were summed.

In the final step to calculate the spectro-temporal SNR, the spectro-temporal power of each target signal was compared to the summed spectro-temporal power of the two masker signals with which it was combined to create a single stimulus. An SNR value was calculated for each T-F unit of the combined three-speaker signal, resulting in a 2-D matrix of local spectro-temporal SNR values (Figure 1, center).

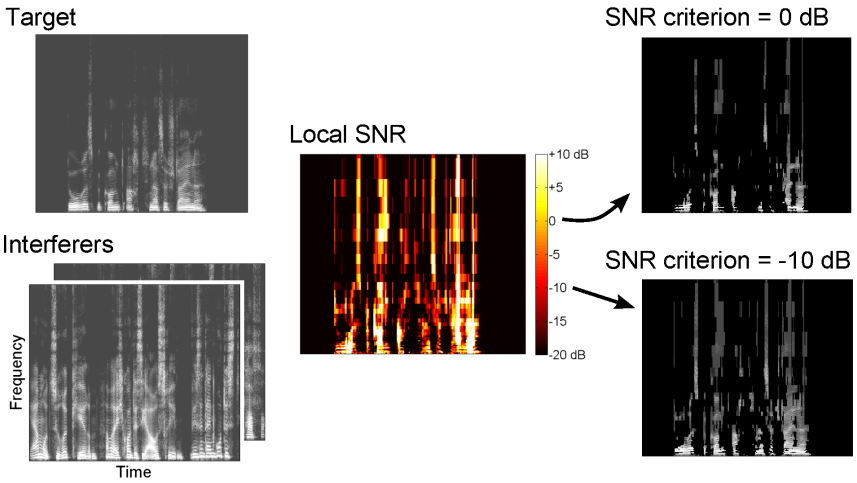


Figure 1. Overview of the stimulus manipulation. Left column: Mean spectrograms of original target and interfering speech. Center: Local SNRs of T-F units. Right column: Spectrograms of target speech after discarding low-SNR T-F units. Examples are shown for SNR criterion values of 0 and -10 dB.

The spectro-temporal representation of the SNR was used to decide which components of the target speech signal would be passed on to the final signal. Any local SNR values exceeding the selection criterion resulted in a value of 1 of the binary mask, and thus in selection of the corresponding spectro-temporal components of the target signal in both ears. Any local SNR values that did not pass the criterion resulted in a value of zero and deselection of the corresponding spectro-temporal components of the target (Figure 1, right column).

A total of 8 different local SNR criteria were included in the experiment based on pilot testing: –10,000, –4, 0, 2, 4, 6, 8 and 10 dB. The criterion of –10,000 dB was chosen to simulate an SNR of –∞, and served as a control condition of unfiltered target speech in which the entire target signal was passed on to the final stimulus.

After addition of the selected spectro-temporal components of the target signal to the complete spectrograms of the two interferer signals, the final signal was reconstructed by calculation of the inverse fast Fourier transform, followed by windowing with square-root Hann windows and overlap-add.

2.3 Procedure

Six normal-hearing participants (4 females, 2 males, aged 22 – 29 years) participated in the experiment. All were native speakers of German.

The subjects listened to the stimuli using Sennheiser HD650 headphones while being seated in a double-walled soundproof booth in front of a PC. The OLSA speech sentence test was run as a closed test and the 5×10 answer alternatives were shown on the screen by means of a graphical user interface. The interface served both to navigate through the experiment and to mark the answers. The participants were instructed to ignore the female speakers, and to mark the perceived words from the male speaker at the frontal position. They were forced to guess upon missed words. The listeners participated in a separate training session in which the target sentences had not been manipulated. The interfering speech and spatial configurations were the same as in the experimental sessions. The training session comprised 160 trials at a global SNR that started from 0 dB and was gradually decreased to the experimental SNR of –8 dB.

During the two experimental sessions that consisted of 160 trials each, all 16 conditions (8 SNR criteria × 2 spatial configurations) were tested interleaved in random order. One test list of 20 sentences, resulting in 100 test words, was used for each condition. Performance was measured as word scoring.

3 Results

Figure 2 shows the percentage of correctly recognized words versus SNR criterion for the six individual listeners. The graphs show a clear spatial release from masking which is reflected by the consistently higher scores for spatially separated speech for all listeners. From Figure 2 it can also be seen that the SNR criterion at which speech intelligibility starts to decrease below the performance in the unfiltered condition (i.e. the leftmost data point at SNR = $-\infty$) differs among listeners. Generally, the transition appears to lie around 0 dB, but whereas some participants showed reduced speech intelligibility already at a criterion of –4 dB, others showed almost unaffected intelligibility up to a criterion of 5 dB SNR in the spatially separated condition.

The mean results are shown in Figure 3, together with the spatial release from masking which equals the difference between the spatially separated and collocated data at each SNR criterion. From this graph it becomes clear that the spatial release from masking remains more or less constant over the complete range of SNR criteria tested. The curves for the separated and collocated curves apparently do not converge.

In order to run statistical analyses on the data, the percentage values were first transformed into rationalized arcsine units (Studebaker, 1985). A two-way repeated-measures ANOVA with factors of spatial condition (collocated or separated) and SNR criterion shows a significant effect of spatial condition [$F(1,5) = 125.68$, $p<0.001$], a significant effect of SNR criterion [$F(7,35) = 34.45$, $p<0.001$], and no significant interaction of spatial condition and SNR criterion. The absence of interaction between spatial condition and SNR criterion confirms the constant spatial release from masking that could be observed in Figure 3.

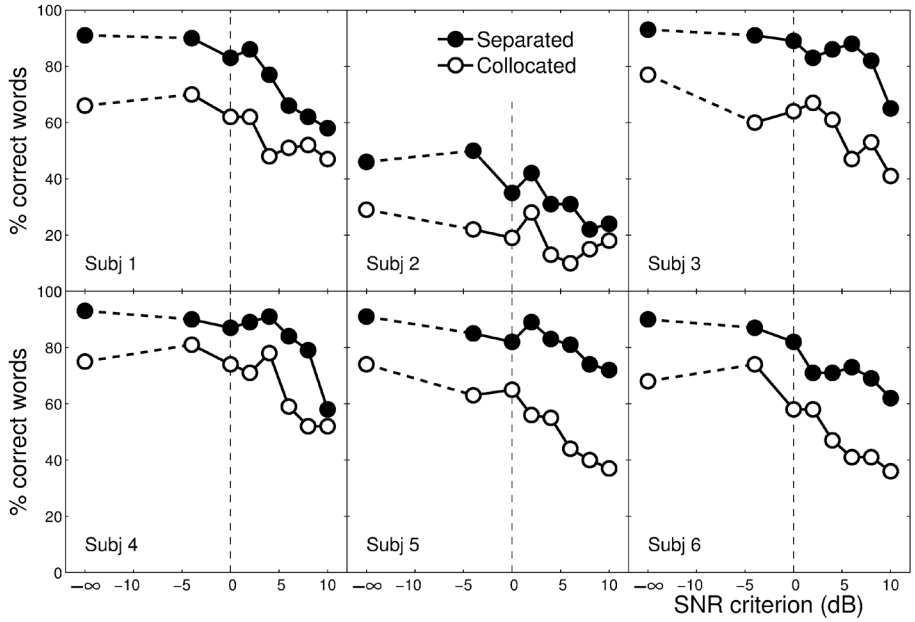


Figure 2. Individual results for six subjects, expressed as percentage correctly identified words versus SNR criterion. Filled circles represent spatially separated speaker configurations; open circles represent collocated configurations.

Simple contrasts comparing scores at each test SNR criterion to the reference criterion at negative infinity, show that all SNR criteria including and exceeding 0 dB resulted in scores different from the reference criterion at the level of $p<0.05$. Only the SNR criterion of –4 dB shows no significant difference from the reference criterion.

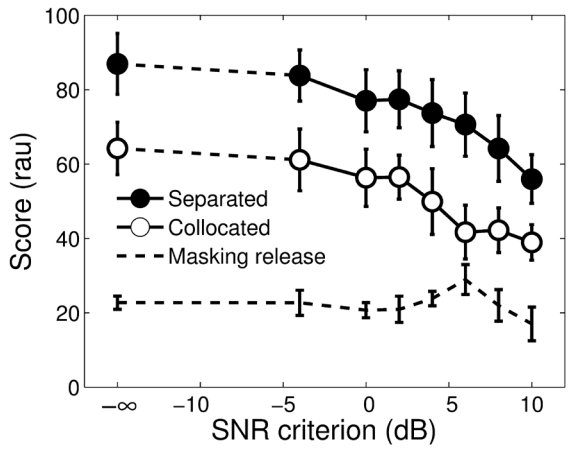


Figure 3. Mean results and spatial release from masking expressed in rau for six subjects. The error bars represent standard errors.

4 Discussion

The aim of this study was to gain more insight into the mechanism leading to spatial release from masking in a multispeaker situation. Specifically, we were interested in assessing whether a BMLD-like detection advantage at the spectro-temporal level could be responsible for access to more target speech elements in a configuration with spatially separated sources, as this could potentially contribute to a higher speech intelligibility.

An experiment in which spectro-temporal regions from the target speech signal were removed according to a variable local SNR criterion gave no indication for any BMLD-like detection advantage. The results of the experiment showed a decrease in speech intelligibility when elements with a local SNR of 0 dB and larger were removed from the target speech, suggesting that speech elements at local SNRs above –4 dB did contribute to speech intelligibility. Interestingly, this was the case for both the spatially collocated and spatially separated speaker configurations. The spatial release from masking appeared to be independent of the extent to which lower SNR components were removed from the target signal.

Our method of applying an SNR-dependent binary mask to the target speech resembles the technique of ideal time-frequency segregation (ITFS) that is known from computational auditory scene analysis studies (e.g., Wang, 2005; Brungart, 2006; Kjems et al., 2009). There are two main differences between the ITFS technique and the technique we have applied in our study. First, we applied the binary mask only to the target speech instead of to the entire speech mixture, leaving the disturbing influence of the interfering speech unchanged over the range of SNR criteria tested. Second, to our knowledge the ITFS technique has only been applied to diotic speech stimuli. Despite these differences the results are comparable in that the studies using ITFS found a decrease in target speech intelligibility for SNR criteria that exceeded the global SNR of the mixture by 5-10 dB. This is comparable to the local SNR value of about 0 dB (i.e. around 10 dB above global SNR) in our study.

An analysis of the local SNRs of all T-F units in 10 unfiltered stimuli (see Figure 4) was performed to determine the proportion of T-F units exceeding a certain SNR level. The analysis reveals that on average only about 20% of the total target T-F units exceeded a local SNR value of –4 dB and about 7% exceeded a value of 10 dB SNR. This corresponds to the proportion of target speech that was retained in the final signal at the SNR criteria mentioned. Note that at the criterion of –4 dB no decrease in speech intelligibility was observed as compared to the condition with intact target speech. This analysis thus demonstrates the sparse character of speech and the fact that only little target speech was needed to achieve a reasonable performance on this experimental task with a restricted speech vocabulary.

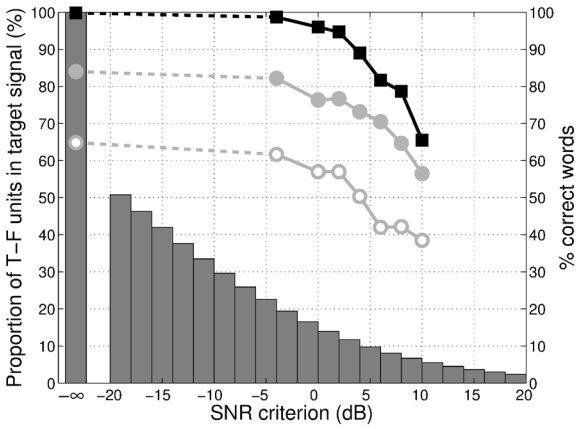


Figure 4. The average proportion of T-F units that were retained in the manipulated target signal for a given SNR criterion are shown by bars (left axis). Black squares show the mean intelligibility of this target speech presented in absence of interferers (right axis). The data from Figure 3 are shown in grey for reference.

The fact that the spatial release from masking remained stable over the range of SNR criteria tested is one of the most important outcomes of this study. It suggests that recovery of sub-threshold speech information by binaural cues does not play a role in the spatial release from masking in a multispeaker situation. Instead, masking release appeared to be sustained even when only high-SNR components of target speech were left in the signal. These supra-threshold components should be detectable monaurally in both the collocated and spatially separated conditions and thus their detection should not rely on a binaural unmasking mechanism as is the case for BMLDs.

An explanation for this spatial release from masking can be provided by directional cues of the stimuli that support a better allocation of speech components to the source of interest. The fact

that speech sources originate from different spatial positions can eliminate some confusion that arises from the simultaneous presence of multiple speech sources (Durlach et al., 2003). This is supported by data (black squares in Figure 4) that show the performance on the same experimental task when the target speech was presented in isolation, but otherwise manipulated identically as before. These data show the inherent intelligibility of the sparse target speech when no interfering speech was present. Especially at the lower range of SNR criteria tested, the data show a similar course as in the presence of interfering speech, but the curves have been shifted towards higher performance with perfect intelligibility in the unfiltered reference condition.

Acknowledgments

This work was supported by the DFG (SFB/TRR31 "The Active Auditory System").

References

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America*, *127*(4), 2479–2497.

Brinkmann, F., Lindau, A., Weinzierl, S., Geissler, G., & van de Par, S. (2013). A high resolutional head-related transfer function database including different orientations of head above the torso. *Fortschritte der Akustik: Tagungsband d. 39. DAGA*, S. 596–599. AIA-DAGA 2013, Merano, Italy: DEGA e.V. Berlin.

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *The Journal of the Acoustical Society of America*, 120(6), 4007–4018.

Durlach, N. I., Mason, C. R., Gerald Kidd, J., Arbogast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003). Note on informational masking (L). *The Journal of the Acoustical Society of America*, *113*(6), 2984–2987.

Giannakopoulos, T. (2010). A method for silence removal and segmentation of speech signals, implemented in Matlab. http://www.mathworks.com/matlabcentral/fileexchange/ 28826-silence-removal-in-speech-signals. Accessed 24 November 2014

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., & Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, *126*(3), 1415–1426.

Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., & Makin, S. J. (2012). Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *The Journal of the Acoustical Society of America*, *131*(1), 218–231.

Levitt, H., & Rabiner, L. R. (1967). Predicting Binaural Gain in Intelligibility and Release from Masking for Speech. *The Journal of the Acoustical Society of America*, *42*(4), 820–829.

Studebaker, G. A. (1985). A „Rationalized" Arcsine Transform. *Journal of Speech, Language, and Hearing Research*, *28*(3), 455–462.

Wagener, K., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie*, *38*, 4–15.

Wan, R., Durlach, N. I., & Colburn, H. S. (2014). Application of a short-time version of the Equalization-Cancellation model to speech intelligibility experiments with speech maskers. *The Journal of the Acoustical Society of America*, *136*(2), 768–776.

Wang, D. (2005). On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. In P. Divenyi (Hrsg.), *Speech Separation by Humans and Machines* (S. 181–197). Springer US.

Chapter 10

On the Contribution of Target Audibility to Performance in Spatialized Speech Mixtures

Virginia Best, Christine R Mason, Jayaganesh Swaminathan, Gerald Kidd, Jr.
Dept. Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA

Kasey M Jakien, Sean D Kampel, Frederick J Gallun
Oregon Health and Science University/VA Portland Health Care System, Portland, OR, USA

Jörg M Buchholz, Helen Glyde
National Acoustic Laboratories, Macquarie University, NSW, Australia

Contact author: VB, ginbest@bu.edu

Abstract

Hearing loss has been shown to reduce speech understanding in spatialized multitalker listening situations, leading to the common belief that spatial processing is disrupted by hearing loss. This paper describes related studies from three laboratories that explored the contribution of reduced target audibility to this deficit. All studies used a stimulus configuration in which a speech target presented from the front was masked by speech maskers presented symmetrically from the sides. Together these studies highlight the importance of adequate stimulus audibility for optimal performance in spatialized speech mixtures and suggest that reduced access to target speech information might explain a substantial portion of the "spatial" deficit observed in listeners with hearing loss.

Keywords: speech, spatial release from masking, audibility, hearing loss

1 Introduction

In the context of speech communication, spatial release from masking (SRM) refers to an improvement in intelligibility when competing sounds are spatially separated from the talker of interest. This improvement can arise as a result of acoustic benefits (such as when the competing sounds are on the opposite side of the head to the target sound and thus are reduced in level at one ear by the "head-shadow") or by effective increases in signal-to-noise ratio resulting from neural processing of interaural cues (so called "masking level differences"). In other cases it appears that the perceived separation of sources drives the advantage by enabling attention to be directed selectively.

In many situations, listeners with sensorineural hearing impairment (HI) demonstrate reduced SRM compared to listeners with normal hearing (NH). This observation commonly leads to the conclusion that spatial processing is disrupted by hearing loss. However, convergent evidence from other kinds of spatial tasks is somewhat lacking. For example, studies that have measured fine discrimination of binaural cues have noted that individual variability is high, and at least some HI listeners can discriminate as well as NH listeners (e.g. Colburn, 1982; Spencer, 2013). In loudspeaker experiments, horizontal localization does not appear to be strongly affected by

hearing loss unless it is highly asymmetric or if the loss is very severe at low frequencies (e.g. Noble et al., 1994). Other studies have tried to relate SRM in multitalker environments to localization ability (Noble et al., 1997; Hawley et al., 1999) or to binaural sensitivity (Strelcyk & Dau, 2009; Spencer, 2013) with mixed results. Finally, it has been observed that SRM is often inversely related to the severity of hearing loss (e.g. Marrone et al., 2008). This raises the question of whether in some cases apparent spatial deficits might be related to reduced audibility in spatialized mixtures.

A popular stimulus paradigm that has been used in recent years consists of a frontally located speech target, and competing speech maskers presented symmetrically from the sides. This configuration was originally implemented to minimize the contribution of long-term head-shadow benefits to SRM (Noble et al., 1997; Marrone et al., 2008) but has since been adopted as a particularly striking case in which the difference between NH and HI listeners is large. This paper describes related studies from three different laboratories that used the "symmetric masker" configuration to explore the interaction between target audibility and performance under these conditions.

2 Part 1

2.1 Motivation

Gallun et al. (2013) found that the effect of hearing loss on separated thresholds was stronger when one target level was used for all listeners (50 dB SPL) compared to when a sensation level (SL) of 30 dB was used (equivalent to a range of 47-72 dB SPL). They speculated that a broadband increase in gain was not sufficient to combat the non-flat hearing losses of their subjects. Thus in their most recent study (Jakien et al., under revision), they performed two experiments. In the first, they presented the same listeners target sentences at either 10 dB SL or 30 dB SL, while in a second experiment they adjusted the audibility within octave frequency bands for each HI listener, to match that of a NH listener.

2.2 Methods

Target and masker stimuli were three male talkers taken from the Coordinate Response Measure corpus. Head-related transfer functions (HRTFs) were used to simulate two situations in which the target sentence was always at 0° azimuth and the maskers were either colocated or positioned at ±45° (Gallun et al., 2013; Xie, 2013). In the first experiment the target sentence was fixed at either 10 dB SL (low SL condition) or 30 dB SL (high SL condition) above each participant's speech reception threshold (SRT) in quiet. The levels of the two masking sentences were adjusted relative to level of the target sentence and were changed in level using a one-up/one-down adaptive tracking algorithm to find threshold target/masker ratio (TMR). In the second experiment the target sentence was adjusted on a band-by-band basis to account for differences in the audiogram across participants. The initial level was set to that of the high SL condition of the first experiment for a listener with 0 dB HL. Target and masking sentences were then filtered into six component waveforms using two-octave-wide bandpass filters with center frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz. The level of each component was adjusted based on the difference between the audiogram of the listener being tested and the audiogram of a comparison listener with 0 dB HL thresholds at each of the six octave frequencies. So, for a listener with a 10 dB HL threshold at 1000 Hz, the level of the 1000 Hz component was amplified by 10 dB. Following adjustment for audibility in this manner, the six waveforms were summed to produce the target or masker waveforms. The levels of the two masking sentences were adjusted relative to level of the target sentence according to a progressive tracking algorithm which has been shown to be comparable to and more efficient than adaptive tracking (Gallun et al., 2013).

Thirty-six listeners participated in both experiments, and an additional 35 participated in just the second experiment. All seventy-one participants had four frequency (500, 1000, 2000, 4000 Hz) average hearing losses (4FAHL) below 37 dB (mean 12.1 dB ± 8.2 dB) and all had fairly symmetrical hearing at 2000 Hz and below (most had differences between the ears of less than 10 dB at all frequencies, and none had differences exceeding 20 dB). Ages of the listeners were between 18 and 77 years (mean of 43.1 years) and there was a moderate ($r^2 = 0.12$) relationship between 4FAHL and age. For simplicity, the listeners have been divided into those with 4FAHLs below 10 dB (n=22 in the first experiment; n=35 in the second experiment) and those with 4FAHLs above 10 dB (n=14 in the first experiment; n=36 in the second experiment).

2.3 Results

In the first experiment (top row of Fig. 1), those with lower 4FAHLs had better thresholds, and thresholds and SRM in both 4FAHL groups improved with an increase in SL. In the second experiment (bottom row of Fig. 1), equating audibility did not remove the performance difference between groups in the spatially separated condition, but did roughly equate SRM.

A series of linear regressions were conducted on the data from the second experiment to analyze relationships between thresholds and 4FAHL, as a function of spatial configuration. In the spatially separated condition, 4FAHL accounted for approximately 8% of the variance in participant performance, while it accounted for 28% of the variance in the colocated condition.

In summary, while increasing SL improved performance for all listeners, neither changing SL nor ensuring audibility for all listeners was able to guarantee equal performance across listeners with various degrees of hearing loss. This suggests that while audibility is clearly an important factor, other factors may impact speech-in-speech intelligibility (e.g. aging, auditory filter width, or comorbidities associated with cognition, working memory and attention). On the other hand, careful equalization of audibility across HI listeners did reduce the variation in SRM observed.

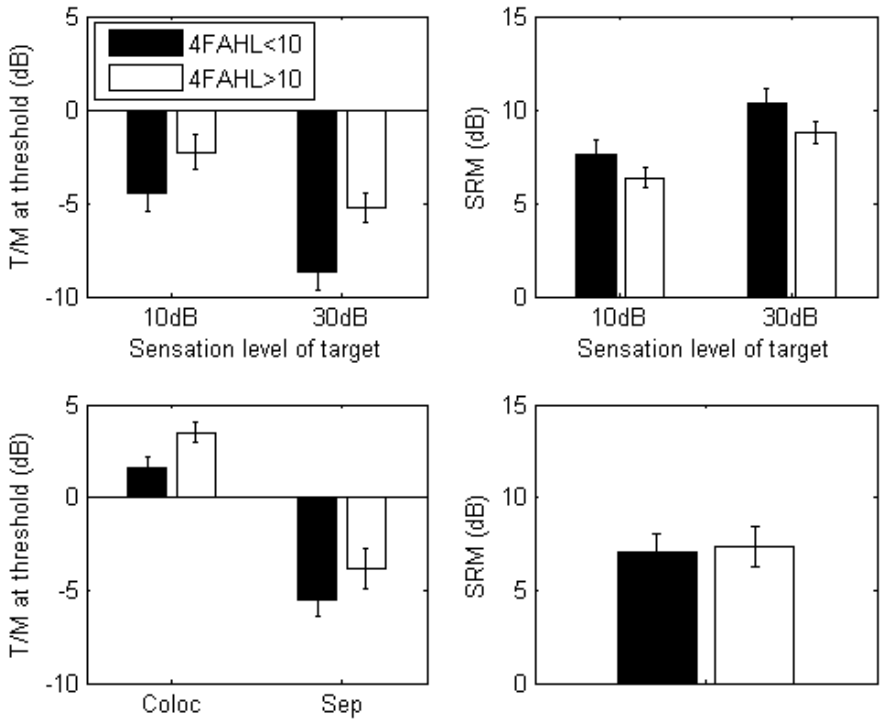


Figure 1. Group-mean TMRs at threshold in the spatially separated conditions (top left) and SRM (top right) as a function of SL. Group mean TMRs at threshold for colocated and separated configurations (bottom left) and SRM (bottom right) in the equal-audibility conditions. Error bars represent two standard errors of the mean.

3 Part 2

3.1 Motivation

Glyde et al. (2013) showed that even with frequency-specific gain applied according to the individual audiogram using the NAL-RP hearing aid prescription, a strong relationship between SRM and hearing status persisted. The authors noted that with the relatively low presentation levels used in their experiment (55 dB SPL masker), the NAL-RP prescription may not have provided sufficient gain especially in the high frequency region. Thus in a follow-up experiment (Glyde et al., under revision), they examined the effect of providing systematically more high-frequency gain than that provided by NAL-RP. They tested HI subjects as well as NH subjects with a simulated hearing loss.

3.2 Methods

The data is compiled from different studies and at least 12 NH (mean age 28.8-33.6 years) and 16 older HI (mean age 68.8-73.1 years) listeners participated in each group. The HI listeners had a moderate, bilaterally symmetric, sloping sensorineural hearing loss with a 4FAHL of 48±5dB.

Subjects were assessed with a Matlab version of the LiSN-S test (Glyde et al., 2013), in which short, meaningful sentences of 5-7 words (e.g.," The brother carried her bag") were presented in an ongoing two-talker background. Target and distractors were realized by the same female talker and target sentences were preceded by a brief tone burst. Using HRTFs, target sentences were presented from 0° azimuth and the distractors from either 0° azimuth (colocated condition: SV0) or ±90° azimuth (spatially separated condition: SV90). The combined distractor level was fixed at 55 dB SPL and the target level was adapted iteratively to determine the TMR at which 50% of the target words were correctly understood. Subjects were seated in an audiometric booth and repeated the target sentences to a conductor.

Stimuli were presented over equalized headphones and for HI listeners had different levels of (linear) amplification applied to systematically vary target/masker audibility: amplification according to NAL-RP, NAL-RP plus 25% of extra gain, and NAL-RP plus 50% of extra gain. Thereby an extra gain of 100% would have restored normal audibility, but was impossible to achieve due to loudness discomfort. Given the sloping hearing loss, an increase in amplification mainly resulted in an increased high-frequency gain and thus in an increase in audible bandwidth. NH subjects were tested at the same audibility levels. This was realized by first applying attenuation filters that mimicked the average audiogram of the HI subjects and then applying the same gains as described above. For the NH subjects the standard NH condition was additionally tested. Details of the processing can be found in Glyde et al. (under revision).

3.3 Results

The left panel of Figure 2 shows 50% thresholds for the spatially separated (SV90) condition and the right panel the corresponding SRM, i.e., the difference in threshold between the spatially colocated (SV0) and separated (SV90) conditions. The spatially separated thresholds clearly improve (decrease) with increasing amplification level for both the NH and HI subjects and are maximal for normal audibility. Applying 50% gain on top of NAL-RP provides an improvement of about 3.4 dB for the NH and 4.8 dB for the HI subject. However, thresholds for the NH subjects are in average 4.8 dB lower than for the HI subjects.

Although not shown here, the thresholds for the colocated conditions are basically independent of amplification level, and for the NH subjects are about 2.5 dB lower than for the HI subjects. Hence, the SRM increases with increasing gain similarly to the spatially separated thresholds, but the overall difference between NH and HI subjects is reduced to about 2.5 dB. Hence, a large proportion of the SRM deficit in the HI (and simulated HI) group could be accounted for by reduced audibility.

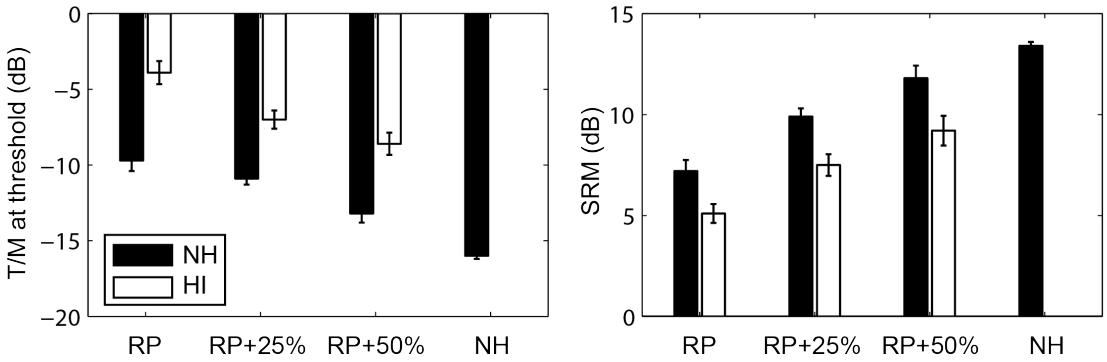


Figure 2. Group-mean TMRs at threshold for the spatially separated (SV90) condition (left panel) and corresponding SRM (right panel). Error bars show standard errors.

4 Part 3

4.1 Motivation

Accounting for audibility effects in speech mixtures is not straightforward. While it is common to measure performance in quiet for the target stimuli used, this does not incorporate the fact that portions of the target are completely masked, which greatly reduces redundancy in the speech signal. Thus a new measure of "masked target audibility" was introduced.

Simple energy-based analyses have been used to quantify the available target in monaural speech mixtures (e.g. the ideal binary mask described by Wang, 2005; ideal time-frequency segregation as explored by Brungart et al., 2006; and the glimpsing model of Cooke, 2006). The basic approach is to identify regions in the time-frequency plane where the target energy exceeds the masker energy. The number of these glimpses is reduced as the SNR decreases, or as more masker talkers are added to the mixture. To define the available glimpses in symmetric binaural mixtures, we simply applied a monaural glimpsing model separately to the two ears. For symmetric listening situations, the glimpses can occur in either ear, and often occur in both ears for a particular time-frequency tile. However the glimpses may not all be above threshold, particularly for listeners with hearing loss. Thus we conducted an experiment in which we presented "glimpsed" stimuli to NH and HI listeners to measure their ability to use the available target information. Performance was compared to natural binaural performance to determine to what extent target audibility/availability can explain performance.

4.2 Methods

Six NH (mean age 23 years) and six HI (mean age 26 years) listeners participated. The HI listeners had a moderate, bilaterally symmetric, sloping sensorineural hearing loss with a 4FAHL of 49±14dB.

Speech materials were taken from a corpus of monosyllabic words (Kidd et al 2008), in which five-word sentences are assembled by selecting a word from each of five categories (e.g., "Sue bought two red toys"). Using HRTFs, one target sentence was presented at 0° azimuth, and two or four different masker sentences were presented at ±90° azimuth, or ±45°/±90° azimuth. All talkers were female, and the target was identified by its first word "Sue". Each masker was fixed in level at 60 dB SPL, and the target was varied in level to set the TMR at one of 5 values. Stimuli for HI listeners had individualized NAL-RP gain applied.

To generate the glimpsed stimuli, an ideal binary mask was applied separately to the two ears of the binaural stimuli using the methods of Wang (2005) and Brungart et al. (2006). Briefly, the signals were analyzed using 128 frequency channels between 80 and 8000 Hz, and 20-ms time windows with 50% overlap. Tiles with target energy exceeding masker energy were assigned a

mask value of 1 and the remaining tiles were assigned a value of 0. The binary mask was then applied to the appropriate ear of the binaural stimulus before resynthesis. As a control condition, the mask was also applied to the target alone.

Stimuli were presented over headphones to the listener who was seated in an audiometric booth fitted with a monitor, keyboard and mouse. Responses were given by selecting five words from a grid presented on the monitor. Stimuli were presented in blocks in which the listening condition (natural, glimpsed mixture, glimpsed target) and number of maskers (two or four) was fixed, but the TMR varied randomly.

4.3 Results

Figure 3 shows 50% thresholds extracted from logistic fits to the data. Performance was better overall with two maskers as compared to four maskers, and for NH than HI listeners. With two maskers, the difference in thresholds between groups was 11 dB in the natural condition, and 9 dB in the glimpsed mixture condition. For four maskers these deficits were 7 dB and 9 dB, respectively. In other words, the group differences present in the natural binaural condition were also largely present when listeners were presented with the good time-frequency glimpses only. In the control condition where the glimpses contained only target energy, the group differences were even larger (12.6 and 10.8 dB). This suggests that the HI deficit is related to the ability to use the available target information and not to difficulties with spatial processing or segregation. Individual performance for natural stimuli was strongly correlated with performance for glimpsed stimuli (r=0.92), again suggesting a common limit on performance in the two conditions.

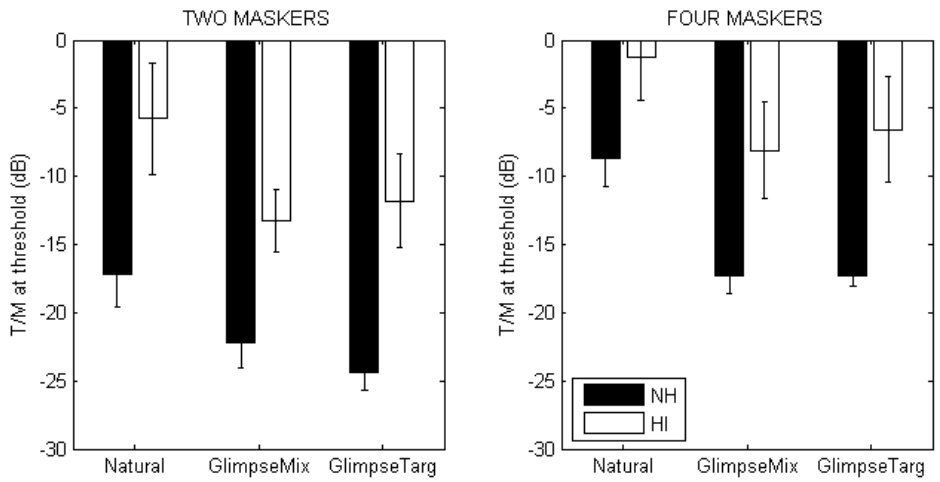


Figure 3. Group-mean TMRs at threshold for the two-masker (left panel) and four-masker (right panel) conditions. Errorbars show standard errors.

5 Conclusions

These studies demonstrate how audibility can affect measures of SRM, and suggest that reduced access to target speech information might in some cases contribute to the "spatial" deficit observed in listeners with hearing loss. This highlights the importance of adequate stimulus audibility for optimal performance in spatialized speech mixtures, although this is not always feasible due to loudness discomfort in hearing aid users or technical limitations within the hearing aids.

Acknowledgments

This work was supported by NIH-NIDCD awards DC04545, DC013286, DC04663, DC00100 and AFOSR award FA9550-12-1-0171 (VB, CRM, JS, GK), DC011828 and the VA RR&D NCRAR (KMJ, SDK, FJG) as well as by the HEARing CRC, established and supported under the Cooperative Research Centres Program – an initiative of the Australian Government (JB, HG).

References

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang D. L. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America,* 120, 4007-4018.

Colburn, H. S. (1982). Binaural interaction and localization with various hearing impairments. *Scandinavian Audiology Supplement*, 15, 27-45.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119, 1562-1573.

Gallun, F. J., Kampel, S. D., Diedesch, A. C., & Jakien, K. M. (2013). Independent impacts of age and hearing loss on spatial release in a complex auditory environment. *Frontiers in Neuroscience*, 252(7), 1-11.

Glyde, H., Cameron, S., Dillon, H., Hickson, L., & Seeto, M. (2013). The effects of hearing impairment and aging on spatial processing. *Ear and Hearing*, 34(1), 15-28.

Hawley, M. L., Litovsky, R. Y., & Colburn, H. S. (1999). Speech intelligibility and localization in a multi-source environment. *Journal of the Acoustical Society of America*, 105, 3436-3448.

Kidd Jr., G., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *Journal of the Acoustical Society of America*, 124, 3793-3802.

Marrone, N., Mason, C. R., & Kidd Jr., G. (2008). The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *Journal of the Acoustical Society of America,* 124, 3064-3075.

Noble, W., Byrne, D., & Lepage, B. (1994). Effects on sound localization of configuration and type of hearing impairment. *Journal of the Acoustical Society of America*, 95, 992-1005.

Noble, W., Byrne, D., & Ter-Host, K. (1997). Auditory localization, detection of spatial separateness, and speech hearing in noise by hearing impaired listeners. *Journal of the Acoustical Society of America*, 102, 2343-2352.

Spencer, N. (2013). Binaural benefit in speech intelligibility with spatially separated speech maskers. PhD Dissertation, Boston University.

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *Journal of the Acoustical Society of America*, 125, 3328-3345.

Wang, D. L. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi (Ed.), *Speech Separation by Humans and Machines* (pp. 181-197). Norwell, MA: Kluwer Academic.

Xie, B. (2013). *Head-Related Transfer Function and Virtual Auditory Display, Second Edition*. Plantation, FL: J. Ross Publishing.

Chapter 11

Optimization of a Spectral Contrast Enhancement Algorithm for Cochlear Implants based on a Vowel Identification Model

Waldo Nogueira[1], Thilo Rode[2], Andreas Büchner[1]

[1] *Dept. of Otolaryngology and Hearing4all, Medical University Hannover, Hannover, Germany*
[2] *HörSys GmbH, Hannover, Germany*

Contact author: WN, nogueiravazquez.waldo@mh-hannover.de

Abstract

Speech intelligibility achieved with cochlear implants (CIs) shows large variability across different users. One reason that can explain this variability is the CI user's individual electrode nerve interface which can impact the spectral resolution they can achieve. Spectral resolution has been reported to be related to vowel and consonant recognition in CI listeners. One measure of spectral resolution is the spectral modulation threshold (SMT), which is defined as the smallest detectable spectral contrast in a stimulus. In this study we hypothesize that an algorithm that improves SMT may improve vowel identification, and consequently produce an improvement in speech understanding for CIs. With this purpose we implemented an algorithm, termed spectral contrast enhancement (SCE) that emphasizes peaks with respect to valleys in the audio spectrum. This algorithm can be configured with a single parameter: the amount of spectral contrast enhancement entitled "SCE factor". We would like to investigate whether the "SCE factor" can be individualized to each CI user. With this purpose we used a vowel identification model to predict the performance produced by the SCE algorithm with different "SCE factors" in a vowel identification task.

In five CI users the new algorithm has been evaluated using a SMT task and a vowel identification task. The tasks were performed for SCE factors of 0 (no enhancement), 2 and 4. In general it seems that increasing the SCE factor produces a decrease in performance in both the SMT threshold and vowel identification.

Keywords: Cochlear Implant, Model, Vowel Identification, Spectral Contrast Enhancement

1 Introduction

Cochlear implants (CIs) can restore the sense of hearing in profound deafened adults and children. CI signal processing strategies have been developed for speech understanding in quiet, such that many post-lingually deafened adults with CIs can recognize 60% to 80% of sentences presented in quiet (Friesen et al., 2001). However, speech intelligibility in noise and music perception, although very variable, remain generally poor for CI listeners.

For example it is still challenging for many CI users to discriminate vowels and phonemes in a closed set identification task without background noise (Sagi et al., 2010; Svirsky et al., 2011). These difficulties might be produced by the limited spectral resolution delivered by CI devices. Spectral resolution may be degraded by the broad electrical fields created in the cochlea when the electrodes are stimulated.

In a recent study, the identification of spectrally smeared vowels and consonants was improved by spectral contrast enhancement (SCE) in a group of 166 normal hearing listeners (Alexander et al., 2011). Spectral contrast is defined as the level difference between peaks and valleys in the spectrum. In CIs, spectral contrast is degraded because of the limited number of stimulation electrodes and overlapping electric fields activating the nervous system through the bony structure of the cochlea. This might reduce the differences in amplitudes between peaks and valleys in the input making it more difficult to locate spectral dominance (i.e., formants) which provide crucial cues to speech intelligibility and instrument identification. Loizou and Poroy (2001) showed that CI users need a higher spectral contrast than normal hearing listeners in vowel identification tasks.

In this study we propose a new sound coding strategy that uses SCE for CI users. The working principle of the coding strategy can affect speech intelligibility. For example "NofM" strategies such as ACE were developed in the 1990s to separate speech signals into M sub-bands and derive envelope information from each band signal. N bands with the largest amplitude are then selected for stimulation (N out of M). One of the consequences here is that the spectral contrast of the spectrum is enhanced, as only the N maxima are retained for stimulation. In this work, we want to investigate whether additional spectral enhancement can provide with improved speech intelligibility.

When designing speech coding strategies, the large variability in speech intelligibility outcomes has to be considered. For example, two sound coding strategies can produce opposite effects in speech performance, even when the CI users are post-locutive adults and have enough experience with their CIs. One possible reason that might explain this variability is the electrode nerve interface of each individual which can impact the spectral resolution they can achieve. Spectral resolution has been reported to be closely related to vowel and consonant recognition in cochlear implant (CI) listeners (Litvak, et al., 2007). One measure of spectral resolution is the spectral modulation threshold (SMT), which is defined as the smallest detectable spectral contrast in the spectral ripple stimulus (Litvak, et al., 2007). In this study we hypothesize that an SCE algorithm may be able to improve SMT and therefore may also be able to improve vowel recognition.

Recently a relatively simple model of vowel identification has been used to predict confusion matrices of CI users. Models of sound perception are not only beneficial in the development of sound coding strategies to prototype the strategy and create hypotheses, but also to give more robustness to the results obtained from an evaluation in CI users. Evaluations with CI users are time consuming and results typically show large variability. In this study we use the same model developed by (Sagi et al., 2010) and (Svirsky et al., 2011) to show the potential benefits of SCE in NofM strategies for CIs.

2 Methods

2.1 The Signal Processing Method: SCE in NofM strategies for CIs

The baseline or reference speech coding strategy is the advanced combinational encoder (ACE, a description of this strategy can be found in Nogueira et al., 2005). The ACE strategy can be summarized in five signal processing blocks: 1) The Fast Fourier Transform (FFT); 2) The envelope detector; 3) The NofM band selection; 4) The loudness growth function (LGF) compression and 5) The channel mapping. The new SCE strategy incorporates a new processing

stage just before the NofM band selection. The goal of this stage is to enhance spectral contrast by attenuating spectral valleys while keeping spectral peaks constant. The amount of spectral contrast enhancement can be controlled by a single parameter termed SCE factor. A more detailed description of the algorithm will be published elsewhere (Nogueira et al., 2015).

2.2 Hardware Implementation

All the stimuli were computed in Matlab© using the ACE and the SCE strategies. The stimuli were output from a standard PC to the Nucleus Cochlear© implant using the Nucleus Interface Communicator (NIC). The Matlab toolbox was used to process the acoustic signals and compute the electrical stimuli delivered to the CI. For each study participant we used their clinical map, i.e., their clinical stimulation rate, comfort and threshold levels, number of maxima and frequency place allocation table. For the experiments presented in the report we used three different configurations of the SCE strategy which only differed in the amount of spectral contrast enhancement applied. The three strategies are denoted by SCE0, SCE2 and SCE4. SCE0 means no spectral enhancement and is exactly the same strategy as the clinical ACE strategy.

2.3 Experiments in Cochlear Implant Users

Participants

Five CI users of the Freedom/System5 system participated in this study. The relevant details for all subjects are presented in Table 1. All the test subjects used the ACE strategy in daily life and all had a good speech performance in quiet.

Each CI user participated in a study to measure SMTs and vowel idenitification performance.

Table 1. Patients details participating in the study

Id.	Age	Side	Cause of deafness	Implant experience in years
1	48	Right	Sudden Hearing Loss	1
2	38	Left	Antibiotics	1.8
3	46	Left	Unknown	7.5
4	25	Right	Ototoxika	7
5	22	Left	Unkwown	3.2

Spectral Modulation Threshold

We used the spectral ripple test presented by Litvak et al., (2007) to estimate the spectral modulation thresholds of each CI user. This task uses a cued two interval, two-alternative, forced choice procedure. In the first interval the standard stimulus was always presented. The standard stimulus had a flat spectrum with bandwidth extending from 350 to 5600 Hz. The signal and the second standard were randomly presented in the other two intervals. Both signals were generated in the frequency domain assuming a sampling rate of 44,100 Hz. The spectral shape of the standard and the signal were generated using the equation:

$$|F(f)| = \begin{cases} 10^{\frac{c}{2}\sin\left(2\pi\left(log_2\left(\frac{f}{350}\right)f_c+\theta_0\right)\right)}, & \text{for } 350 < f < 5600 \\ 0, & \text{otherwise} \end{cases}$$

where F is the amplitude of a bin with center frequency f (in Hertz), f_c is the spectral modulation frequency (in cycles/octave), and θ_0 is the starting phase. Next, noise was added to the phase of each bin prior to computing the inverse Fourier transform. The standard was generated using a

spectral contrast *c* equal to 0. The amplitude of each stimulus was adjusted to an overall level of 60 dB sound pressure level (SPL). Independent noise stimuli were presented on each observation interval. The stimulus duration was 400 ms. A 400 ms pause was used between the stimuli.

Thresholds were estimated using an adaptive psychophysical procedure employing 60 trials. The signal contrast level was reduced after three consecutive correct responses and increased after a single incorrect response. Initially the contrast was varied in a step size of 2 dB, which was reduced to 0.5 dB after three reversals in the adaptive track (Levitt, 1971). Threshold for the run was computed as the average modulation depth corresponding to the last even number of reversals, excluding the first three. Using the above procedure, modulation detection thresholds were obtained for the modulation frequency of 0.5 cycles/octave which is the one that correlates best with vowel identification (Litvak et al., 2007).

Vowel Identification Task

Speech understanding was assessed using a vowel identification task. Vowel stimuli consisted of eight long vowels ‘baat', 'baeaet', 'beet', 'biit', 'boeoet', 'boot', 'bueuet', 'buut'. All vowels had a very similar duration of around 180 ms. The stimuli were uttered by a woman. An 8-alternative forced choice task procedure 8-AFC was created where 2 and 4 repetitions of each vowel were used for training and testing respectively. The vowels were presented at the same 60 dB SPL level as the spectral ripples with a loudness roving of +/- 1.5 dB.

The standard Multidimensional Phoneme Identification Model

We used a model of vowel identification to select the amount of spectral contrast enhancement SCE factor. The model is based on the multidimensional phoneme identification (MPI) model (Sagi et al., 2010; Svirsky et al., 2011). A basic block diagram of the model is presented in Figure 1.

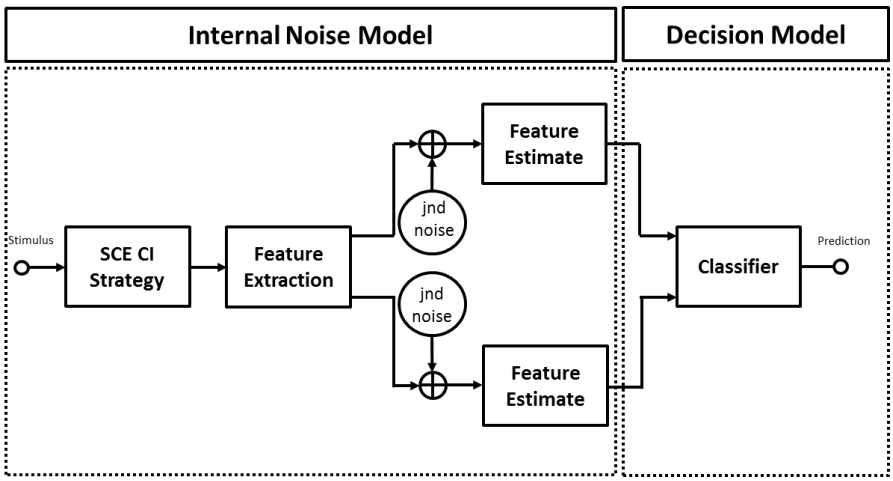


Figure 1. Two dimensional implementation of the MPI model from Sagi et al. (2010) and Svirsky et al. (2011)

The model estimates relevant features from electrodograms generated by the CI sound processor. Because we are modeling a vowel identification task it makes sense to extract features related to formant frequencies. Moreover, because we are analyzing the effect of enhancing spectral contrast it seems logical to use spectral contrast features between formants. In this study the number of features was limited to two formants and therefore the MPI model is two dimensional. Next, the MPI model adds noise to the features. The variance of the noise is set based on the individual abilities of a CI user to perceive the features extracted from the electrodogram. In our implementation we used the results of the SMT task to set the noise

variance. The obtained jnd from the SMT task was scaled between 0.001 and 0.5. This number was used as the variance of the Gaussian noise used in the MPI model and for this reason is termed jnd noise in Figure 1.

3 Results

3.1 Results from the MPI model

The MPI model has been used to model the performance of the SCE strategy for five virtual CI users. All virtual CI user differs from each other in their most comfortable and threshold levels used in the map. That means that the speech processor of each virtual CI user will generate different electrodograms for the same vowels. Next, formant features were extracted from the electrodograms based on the spectral contrast between formants 1 and 2. Noise was added to the spectral contrast features (jnd noise). Three amounts of noise were added 0.01, 10 and 50 percent of the magnitude of the formants extracted. Figure 2 presents the results predicted by the model in percentage of correct vowels identified for five different SCE factors (0, 0.5, 1, 2 and 4).

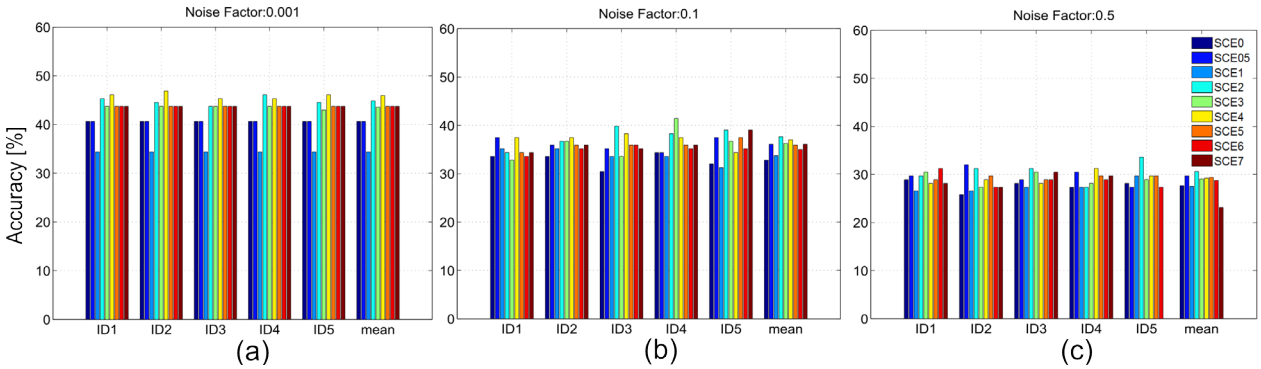


Figure 2. Percentage of correctly identified vowels predicted by the MPI model for 5 virtual CI users using 5 different amounts of SCE (0, 0.5, 1, 2 and 4) and for different amounts of internal noise (Noise Factors 0.001, 0.1 and 0.5).

From the modelling results it can be observed that maximum performance is achieved for SCE factors 2 and 4. The other interesting aspect is that there is no difference in performance across the five different virtual CI users for "Noise Factor" 0.001 (Figure 2a). That means that the feature extraction (i.e. the extraction of the formants) is robust to the different electrodograms of each virtual CI user. Differences in performance between the five virtual CI users can only be observed for "Noise Factor" 0.1 and 0.5, meaning that the "jnd noise" is the only parameter explaining the differences.

From the MPI modelling results we decided to use SCE factors 2 and 4 (equivalent to increasing the original spectral contrast of the spectrum by 3 and by 5 in a dB scale) to be investigated in CI users.

3.2 Results Vowel Identification in CI users

Figure 3 presents the averaged results of the vowel identification task for the five subjects participating in this study.

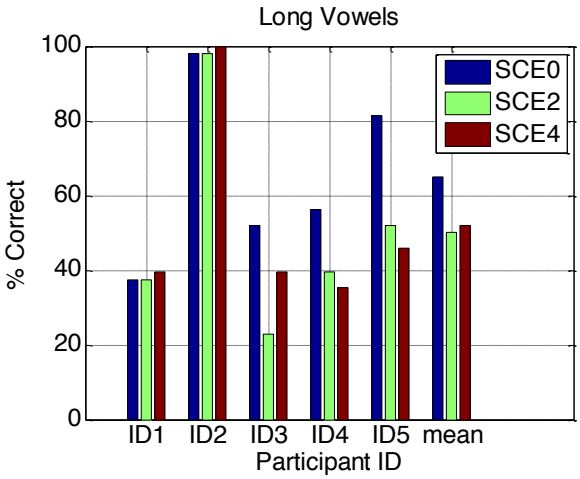


Figure 3. Results of the vowel identification task for the three strategies (SCE0, SCE2 and SCE4) for 5 CI users and averaged as % correct.

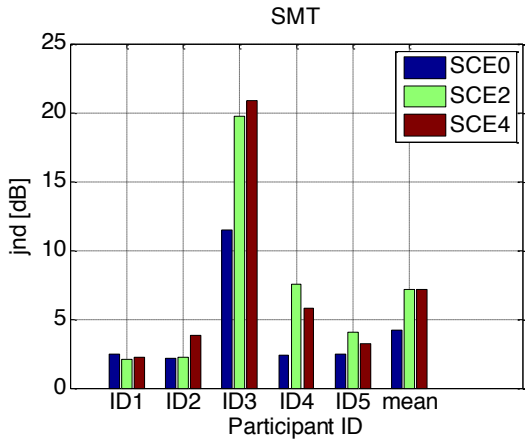


Figure 4. Results of the spectral modulation threshold task for the three strategies (SCE0, SCE2 and SCE4) given as just noticeable differences (jnd SMT) in decibels. The lower is the jnd the better is the result.

3.3 Results Spectral Modulation Threshold in CI users

Figure 4 presents the individual and averaged results for the spectral ripple task.

Unexpectedly, additional spectral contrast, which in turn increases the spectral modulation depth, could no produce an improvement in jnd SMT.

3.4 Correlation between Spectral Modulation Threshold and Vowel Identification

An important question for the analysis was whether the results obtained from the spectral ripple task could be used to predict the outcome of the vowel identification task. This can be seen in the left plot of Figure 5 using an SCE factor of 0. Probably because of the low number of participants only a relatively weak correlation between the two measures was observed.

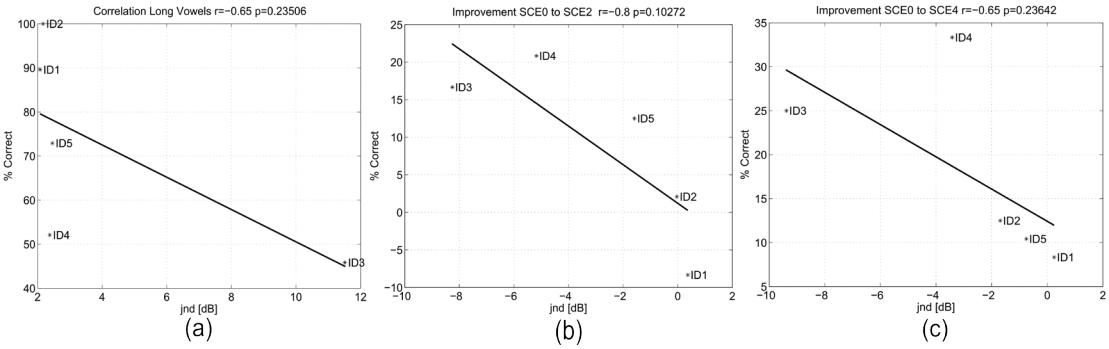


Figure 5. a) Correlation between vowel identification performance and jnd SMT in dB b) Correlation between the difference (SCE0–SCE2) in vowel identification performance and the difference in jnd SMT (SCE0-SCE2) c) Correlation between the difference (SCE0–SCE4) in vowel identification performance and the difference in jnd SMT (SCE0-SCE4).

In the same manner, the middle and right plots in Figure 5 show the relationship between the improvements of the two tasks comparing the results using an SCE factor 0 to those using SCE factors 2 and 4 respectively. Again, the correlation observed is weak but still a trend for the relationship can be seen. It seems that for the SCE factors used the decline in performance in the SMT is somewhat connected to a decline in performance in the vowel identification task. It remains unclear if an increase of the number of participants would confirm this trend.

4 Discussion

A new sound coding strategy that enhances spectral contrast has been designed. The amount of spectral contrast enhancement can be controlled by a single parameter termed SCE factor.

A model of vowel identification has been used to investigate the effect of SCE on vowel identification. The model predicts that increasing the amount of SCE increases vowel identification accuracy. Based on these results we decided to use SCE factors 2 and 4 (equivalent to increasing the original spectral contrast of the spectrum by 3 and by 5 in a dB scale).

The new SCE strategy has been evaluated in CI users. Results from a vowel identification task and a SMT task in five CI users show differences in vowel identification scores for different SCE factors. In general, it seems that SCE produces a decremental effect in spectral modulation detection and vowel identification in CI users. These results are contrary to the model predictions. Previous studies in the literature give reasons to believe that spectral contrast enhancement would result in a benefit for the chosen tasks. It is possible that spectral valleys are attenuated too much and relevant information required by the CI users to understand speech is lost. These effects are not taken into account by the MPI model, and this could explain the contradictory results between experiments and CI users and modelling results. Still, it is possible that the SCE factors selected where too high, for this reason we think that a follow-up study should investigate whether lower amounts SCE can provide improvements in CI users.

Acknowledgments

The authors would like to thank the subjects who have participated in the experiments and the two anonymous reviewers for their comments on different versions of this manuscript. This work was supported by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all” and Cochlear.

References

Alexander, J. M., Jenison, R. L., & Kluender, K. R. (2011). Real-time contrast enhancement to improve speech recognition. *PloS One*, *6*(9), e24630. doi:10.1371/journal.pone.0024630

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, *110*(2), 1150. doi:10.1121/1.1381538

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2), Suppl 2:467+. doi:10.1121/1.1912375

Litvak, L. M., Spahr, A. J., Saoji, A. a, & Fridman, G. Y. (2007). Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners. *The Journal of the Acoustical Society of America*, *122*(2), 982–91. doi:10.1121/1.2749413

Loizou, P. C., & Poroy, O. (2001). Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners. *The Journal of the Acoustical Society of America*, *110*(3 Pt 1), 1619–1627. doi:10.1121/1.1388004

Nogueira, W., Büchner, A., Lenarz, T., & Edler, B. (2005). A psychoacoustic "NofM"-type speech coding strategy for cochlear implants. *Eurasip Journal on Applied Signal Processing*, *2005*, 3044–3059. doi:10.1155/ASP.2005.3044

Nogueira, T. Rode, A. Büchner. (2015). Spectral contrast enhancement improves speech intelligibility for cochlear implants. Submitted to: *Journal of the Acoustical Society of America*, 2015

Sagi, E., Meyer, T. a, Kaiser, A. R., Teoh, S. W., & Svirsky, M. a. (2010). A mathematical model of vowel identification by users of cochlear implants. *The Journal of the Acoustical Society of America*, *127*(2), 1069–83. doi:10.1121/1.3277215

Svirsky, M. a, Sagi, E., Meyer, T. a, Kaiser, A. R., & Teoh, S. W. (2011). A mathematical model of medial consonant identification by cochlear implant users. *The Journal of the Acoustical Society of America*, *129*(4), 2191–200. doi:10.1121/1.3531806

Chapter 12

Roles of the Contralateral Efferent Reflex in Hearing demonstrated with Cochlear Implants

Enrique A. Lopez-Poveda, Almudena Eustaquio-Martín
INCYL, IBSAL, Dept. of Surgery, Medical School, University of Salamanca, Salamanca, Spain

Joshua S. Stohl, Robert D. Wolford
USA Laboratory, MED-EL GmbH, Innsbruck, Austria

Reinhold Schatzer
Institute of Mechatronics, University of Innsbruck, Innsbruck, Austria

Blake S. Wilson
Duke University, Durham, North Carolina, USA

Contact author: EAL-P, ealopezpoveda@usal.es

Abstract

Our two ears do not function as fixed and independent sound receptors; their functioning is linked and dynamically adjusted via the contralateral medial olivocochlear efferent reflex (MOCR). The MOCR possibly facilitates speech recognition in noisy environments. Such a role, however, is yet to be demonstrated because selective deactivation of the reflex during natural acoustic listening has not been possible for human subjects up until now. Here, we propose that this and other roles of the MOCR may be elucidated using the unique stimulus controls provided by cochlear implants (CIs). Pairs of sound processors were constructed to mimic or not mimic the effects of the contralateral MOCR with CIs. For the non-mimicking condition (STD strategy), the two processors in a pair functioned independently of each other. When configured to mimic the effects of the MOCR (MOC strategy), however, the two processors communicated with each other and the amount of compression in a given frequency channel of each processor in the pair decreased with increases in the output energy from the contralateral processor. The analysis of output signals from the STD and MOC strategies suggests that in natural binaural listening, the MOCR possibly causes a small reduction of audibility but enhances frequency-specific inter-aural level differences, and the segregation of spatially non-overlapping sound sources. The proposed MOC strategy could improve CI and hearing-aid performance.

Keywords: olivocochlear reflex, speech intelligibility, spatial release from masking, unmasking, dynamic compression, auditory implants, hearing aids, speech encoding, sound processor

1 Introduction

The central nervous system can control sound coding in the cochlea via medial olivocochlear (MOC) efferents (Guinan, 2006). MOC efferents act upon outer hair cells, inhibiting the

mechanical gain in the cochlea for low- and moderate-intensity sounds (Cooper and Guinan, 2006) and may be activated involuntarily (i.e., in a reflexive manner) by ipsilateral and/or contralateral sounds (Guinan, 2006). This suggests that auditory sensitivity and dynamic range could be dynamically and adaptively varying during natural, binaural listening depending on the state of activation of MOC efferents.

The roles of the MOC reflex (MOCR) in hearing remain controversial (Guinan, 2006; Guinan, 2010). The MOCR may be essential for normal development of cochlear active mechanical processes (Walsh et al., 1998), and/or to minimize deleterious effects of noise exposure on cochlear function (Maison et al., 2013). A more widely accepted role is that the MOCR possibly facilitates understanding speech in noisy environments. This idea is based on the fact that MOC efferent activation restores the dynamic range of auditory nerve fibre responses in noisy backgrounds to values observed in quiet (Fig. 5 in Guinan 2006), something that probably improves the neural coding of speech embedded in noise (Brown et al., 2010; Chintanpalli et al., 2012; Clark et al., 2012). The evidence in support for this unmasking role of the MOCR during natural listening is, however, still indirect (Kim et al., 2006). A direct demonstration would require being able to selectively deactivate the MOCR while keeping afferent auditory nerve fibres functional. To our knowledge, such an experiment has not been possible for human subjects up until now. We argue that the experiment can be done in an approximate manner using cochlear implants (CIs).

Cochlear implants can enable useful hearing for severely to profoundly deaf persons by bypassing the cochlea via direct electrical stimulation of the auditory nerve. The sound processor in a CI (Figure 1) typically includes an instantaneous back-end compressor in each frequency channel of processing to map the wide dynamic range of sounds in the acoustic environment into the relatively narrow dynamic range of electrically evoked hearing (Wilson et al., 1991). The compressor in a CI serves the same function as the cochlear mechanical compression in a healthy inner ear (Robles and Ruggero, 2001). In the fitting of modern CIs, the amount and endpoints of the compression are adjusted for each frequency channel and its associated intracochlear electrode(s). The compression once set is fixed. Cochlear implant users thus lack MOCR effects, but these effects can be reinstated and thus the role(s) of the MOCR in hearing can be assessed by using dynamic rather than fixed compression functions. For example, contralateral MOCR effects could be reinstated by making compression in a given frequency channel vary dynamically depending upon an appropriate control signal from a contralateral processor (Figure 1). Ipsilateral MOCR effects could be reinstated similarly, except that the control signal would originate from the ipsilateral processor. Using this approach, the CI offers the possibility to mimic (with dynamic compression) or not mimic (with fixed compression) the effects of the MOCR. Such activation or deactivation of MOCR effects is not possible for subjects who have normal hearing.

Our long-term goal is to mimic MOCR effects with CIs. This would be useful not only to explore the roles of the MOCR in acoustic hearing but to also reinstate the potential benefits of the MOCR to CI users. The present study focuses on (1) describing a bilateral CI sound processing strategy inspired by the *contralateral* MOCR; (2) demonstrating, using CI processor output-signal simulations, some possible benefits of the contralateral MOCR in acoustic hearing.

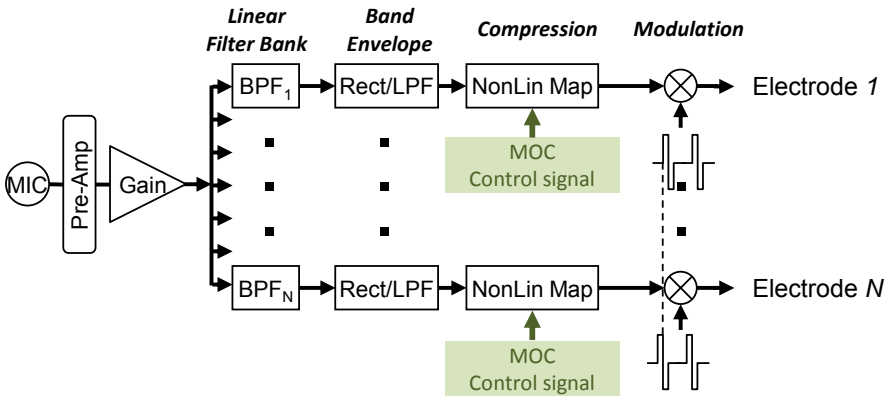


Figure 1. Envisaged way of incorporating MOC efferent control to a standard pulsatile CI sound processor. The elements of the standard sound processor are depicted in black (Wilson et al., 1991) and the MOC efferent control in green. See the main text for details.

2 Methods

2.1 A bilateral CI sound processor inspired by the contralateral MOCR

To assess the potential effects of the contralateral MOCR in hearing using CIs, pairs of sound processors (one per ear) were implemented where the amount of compression in a given frequency channel of each processor in the pair was either constant (STD strategy) or varied dynamically in time in a manner inspired by the contralateral MOCR (MOC strategy).

The two processors in each pair were as shown in Figure 1. They included a high-pass pre-emphasis filter (first-order Butterworth filter with a 3-dB cutoff frequency of 1.2 kHz); a bank of 12, sixth-order Butterworth band-pass filters whose 3-dB cut-off frequencies followed a modified logarithmic distribution between 100 and 8500 Hz; envelope extraction via full-wave rectification (Rect.) and low-pass filtering (LPF) (fourth-order Butterworth low-pass filter with a 3-dB cut-off frequency of 400 Hz); and a back-end compression function.

The back-end compression function was as follows (Boyd, 2006):

$$y=\log(1+c\cdot x)/\log(1+c) \quad , \qquad (1)$$

where x and y are the input and output amplitudes to/from the compressor, respectively, both of them assumed to be within the interval [0,1]; and c is a parameter that determines the amount of compression. In the STD strategy, c was constant ($c = c_{max}$) and identical at the two ears. In the MOC strategy, by contrast, c varied dynamically in time within the range $[c_{max}, c_{min}]$, where c_{max} and c_{min} produce the most and least amount of compression, respectively. For the natural MOCR, the more intense the contralateral stimulus, the greater the amount of efferent inhibition and the more linear the cochlear mechanical response (Hood et al., 1996). On the other hand, it seems reasonable that the amount of MOCR inhibition depends on the magnitude of the output from the cochlea rather than on the level of the acoustic stimulus. Inspired by this, we assumed that the instantaneous value of c for any given frequency channel in an MOC processor was inversely related to the output energy, E, from the corresponding channel in the contralateral MOC processor. In other words, the greater the output energy from a given frequency channel in the left-ear processor, the less the compression in the corresponding frequency channel of the right-ear processor, and vice versa (contralateral 'on-frequency' inhibition). We assumed a

relationship between c and E such that in absence of contralateral energy, the two MOC processors in the pair behaved as a STD ones (i.e., $c = c_{max}$ for $E = 0$).

Inspired by the exponential time courses of activation and deactivation of the contralateral MOCR (Backus and Guinan, 2006), the instantaneous output energy from the contralateral processor, E, was calculated as the root-mean-square (RMS) output amplitude integrated over a preceding exponentially decaying time window with two time constants.

2.2 Evaluation

The compressed output envelopes from the STD and MOC strategies were compared for different signals in different types and amounts of noise or interferers (see below). Stimulus levels were expressed as RMS amplitudes in full-scale decibels (dBFS), i.e., dB re the RMS amplitude of a 1-kHz sinusoid with a peak amplitude of 1 (i.e., dB re 0.71).

The effects of contralateral inhibition of compression in the MOC strategy were expected to be different depending upon the relative stimulus levels and frequency spectra at each ear. For this reason, and to assess performance in binaural, free-field scenarios, all stimuli were filtered through KEMAR head-related transfer functions (HRTFs) (Gardner and Martin, 1995) prior to their processing through the MOC or STD strategies. HRTFs were applied as 512-point finite impulse response filters. STD and MOC performance was compared assuming sound sources at eye level (0° of elevation).

For the present evaluations, c_{max} was set to 1000, a value typically used in clinical CI sound processor, and c_{min} to 1. The energy integration time constants were 4 and 29 ms. These parameter values were selected *ad hoc* to facilitate visualising the effects of contralateral inhibition of compression in the MOC strategy and need not match natural MOCR characteristics or improve CI-user performance (see the Discussion).

All processors were implemented digitally in the time domain, in Matlab™ and evaluated using a sampling frequency of 20 kHz.

3 Results

3.1 The MOC processor enhances within-channel inter-aural level differences

The compressed envelopes at the output from the STD and MOC processors were first compared for a 1-kHz pure tone presented binaurally from a sound source in front of the left ear. The tone had a duration of 50 ms (including 10-ms cosine squared onset and offset ramps) and its level was –20 dBFS. Figure 2A shows the output envelopes for channel #6. Given that the stimulus was binaural and filtered through appropriate HRTFs, there was an acoustic inter-aural level difference (ILD), hence the greater amplitude for the left (blue traces) than for the right ear (red traces). Strikingly, however, the inter-aural output difference (IOD, black arrows in Figure 2A) was greater for the MOC than for the STD strategy. This is because the gain (or compression) in the each of the two MOC processors was inhibited in direct proportion to the output energy from the corresponding frequency channel in the contralateral processor. Because the stimulus level was higher on the left ear, the output amplitude was also higher for the left- than for the right-ear MOC processor. Hence, the left-ear MOC processor inhibited the right-ear MOC processor more than the other way around.

Of course, the gains of the left-ear MOC processor were also inhibited by the right-ear processor to some extent. As a result, the output amplitude in the left-ear processor was slightly lower for the MOC than for the STD strategy (green arrow in Figure 2A). The amount of this inhibition was, however, smaller than for the right-ear MOC processor because (1) the stimulus level was lower on the right ear, and (2) the gain in the right-ear processor was further inhibited by the higher left-ear processor output. In other words, the asymmetry in the amount of contralateral

inhibition of gain (or compression) was enhanced because it depended on the processors' *output* rather on their input amplitudes.

Figure 2B shows that the compression parameter, *c*, in channel #6 of the MOC processors was dynamically varying, hence the IOD was relatively smaller at the stimulus onset and increased gradually in time until it reached a plateau effect. The figure also illustrates the time course of this effect. Note that at time = 0 s, compression was identical for the right-ear and left-ear MOC processors, and equal to the compression in the STD processors (c = 1000). Then, as the output amplitude from the left-ear processor increased, compression in the right-ear processor decreased (i.e., Eq. (1) became more linear). This decreased the output amplitude from the right-ear processor even further, which in turn reduced the inhibitory effect of the right-ear processor on the left-ear processor.

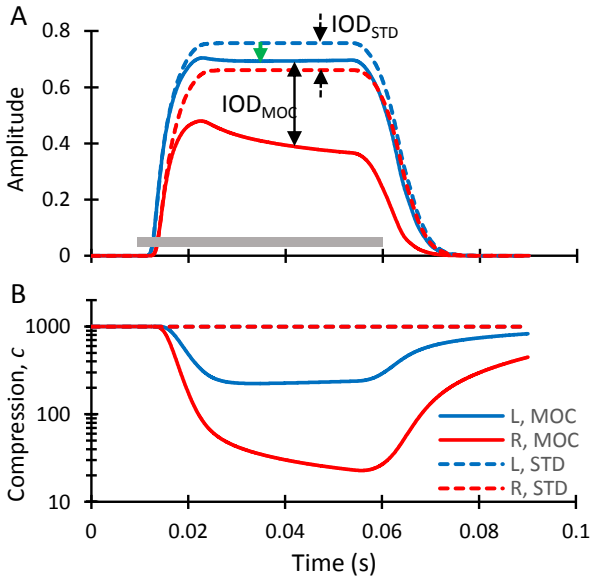


Figure 2. Compressed output envelopes (**A**) and compression parameter (**B**) for frequency channel #6 of the STD and MOC processors in response to a 1-kHz pure tone presented binaurally from a sound source in front of the left ear. Red and blue traces depict signals at the left (L) and right (R) ears, respectively. The grey horizontal line depicts the time period when the stimulus was on. Note the overlap between the dashed red and blue traces in the bottom panel.

Unlike the MOC processor, whose output amplitudes and compression varied dynamically depending on the contralateral output energy, compression in the two STD processors was fixed (i.e., c was constant as depicted by the dashed lines in Figure 2B). As a result, IOD_{STD} was a compressed version of the acoustic ILD and smaller than IOD_{MOC}. In summary, the MOC processor enhanced the within-channel ILDs and presumably conveyed a better lateralized signal compared to the STD processor.

3.2 The MOC processor enhances the spatial segregation of simultaneous sounds

Spatial location is a powerful cue for auditory streaming (Bregman, 2001). This section is devoted to showing that the MOC processor can enhance the spatial segregation of speech sounds in a multi-talker situation.

Figure 3 shows example 'electrodograms' (i.e., graphical representations of processors' output amplitudes as a function of time and frequency channel number) for STD and MOC strategies for two disyllabic Spanish words uttered simultaneously in the free field by speakers located on either side of the head: the word *'diga'* was uttered on the left side, and the word '*sastre*' was uttered on the right side of head. In this example, the two words were actually uttered by the same female speaker and were presented at the same level (–20 dBFS). Approximate spectrograms for the two words are shown in Figure 3A and Figure 3B as the output from the processors' filter banks (BPF in Figure 1). Note that up to this stage, processing was identical and linear for the STD and MOC strategies.

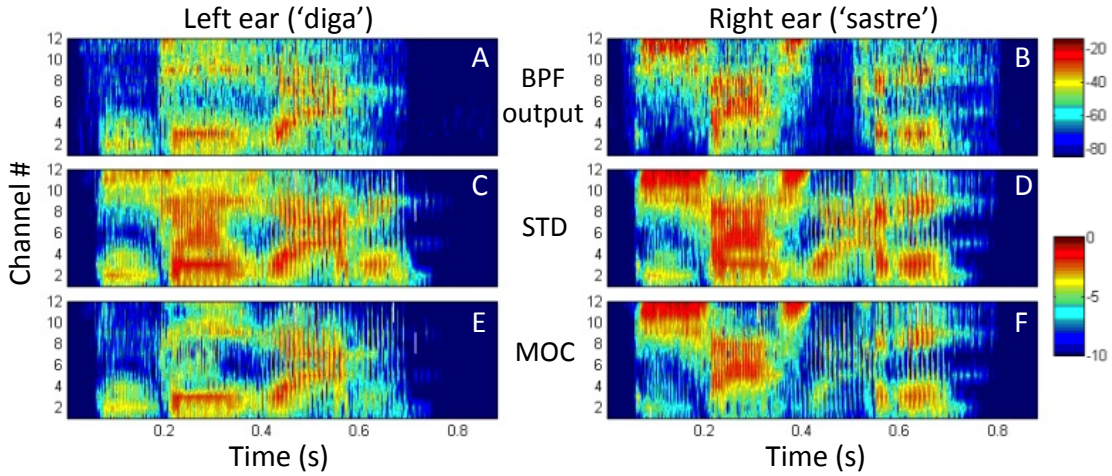


Figure 3. A comparison of left- and right-ear 'electrodograms' for the STD and MOC strategies. The stimulus mimicked a free-field multi-talker condition where a speaker in front of the left ear uttered the Spanish word '*diga*' while a second speaker in front of the right ear uttered the word '*sastre*'. **A, B.** Output from the processors linear filter banks (BPF in Figure 1) for each separate word. **C, D.** Electrodograms for the left- and right-ear STD processors. **E, F.** Electrodograms for the left- and right-ear MOC processors. Colour depicts output amplitudes in dBFS.

The overlap between the words' most significant features was less for the MOC (Figure 3E and Figure 3F) than for the STD strategy (Figure 3C and Figure 3D). For the STD strategy (middle row), significant features of the two words were present in the outputs of the left- and right-ear processors. By contrast, one can almost 'see' that the word *diga*, which was uttered on the left side of the head, was encoded mostly in the output of the left MOC processor, and the word *sastre*, which was uttered on the right side of the head, was encoded mostly in the right MOC processor output, with little interference between the two words at either ear. This suggests that the MOC processor enhances the lateralization of speech, and possibly spatial segregation, in situations with multiple spatially non-overlapping speakers. Although not shown, something similar occurred for speech presented in competition with noise. This suggests that the MOC processor may (1) enhance speech intelligibility in noise or in 'cocktail party' situations, and (2) reduce listening effort.

4 Discussion and Conclusions

Our main motivation was to gain insight into the roles of the MOCR in acoustic hearing by using the unique stimulus controls provided by CIs. As a first step, we have proposed a bilateral CI sound coding strategy inspired by the contralateral MOCR and shown that, compared to using functionally independent sound processors on each ear, our strategy can enhance ILDs on a within-channel basis (Figure 2) and the segregation of spatially distant sound sources (Figure 3). An extended analysis (not shown here due to space restrictions) revealed that it can also enhance within-channel amplitude modulations (a natural consequence from using less compression) and improve the speech-to-noise ratio in the speech ear for spatially non-overlapping speech and noise sources. The proposed MOC processor might thus improve the intelligibility of speech in noise, and the spatial segregation of concurrent sound sources, while reducing listening effort. These effects and their possible benefits have been inferred from analyses of MOC-processor output envelopes for a limited number of spatial conditions. Though promising, further research is necessary to confirm that they hold for other spatial conditions and for behavioural hearing tasks.

The proposed MOC strategy is inspired by the contralateral MOCR but is not an accurate model of it. So far, we have disregarded that contralateral MOCR inhibition is probably stronger for apical than for basal cochlear regions (Lilaonitkul and Guinan, 2009a; Aguilar et al., 2013) and that there is a half-octave shift in frequency between the inhibited cochlear region and the spectrum of contralateral MOCR elicitor sound (Lilaonitkul and Guinan, 2009b). In addition, the present MOC-strategy parameters were selected *ad hoc* to facilitate visualization of the reported effects rather than to realistically account for the time course (Backus and Guinan, 2006) and magnitude (Cooper and Guinan, 2006) of contralateral MOCR gain inhibition. Our next steps include assessing the potential benefits of more accurate MOC-strategy implementations and parameters.

Despite these inaccuracies, it is tempting to conjecture that the natural contralateral MOCR can produce effects and benefits for acoustic hearing similar to those reported here for the MOC processor. Indeed, Figure 2 is consistent with the previously conjectured role of the contralateral MOCR as a 'cocktail party' processor (Kim et al., 2006). The present analysis shows that this role manifests during 'free-field' listening and results from a combination of head-shadow frequency-specific ILDs with frequency-specific contralateral MOCR inhibition of cochlear gain. This and the other possible roles of the MOCR described here would complement its more widely accepted role of improving the neural encoding of speech in noisy environments by restoring the effective dynamic range of auditory nerve fibre responses (Guinan, 2006). The latter effect is neural and cannot show up in the MOC-processor output signals used in the present analyses.

The potential benefits attributed here to the contralateral MOCR are unavailable to current CI users and this might explain part of their impaired performance understanding speech in challenging environments. The present results suggest that the MOC processor, or a version of it, might improve CI performance. The MOC strategy may also be adapted to bilateral hearing aids.

Acknowledgments

We thank Enzo L. Aguilar, Peter T. Johannesen, Peter Nopp, Patricia Pérez-González, and Peter Schleich for comments. This work was supported by the Spanish Ministry of Economy and Competitiveness (grants BFU2009-07909 and BFU2012-39544-C02-01) to EAL-P (principal investigator) and by MED-EL GmbH.

References

Aguilar E, Eustaquio-Martin A, Lopez-Poveda EA (2013) Contralateral efferent reflex effects on threshold and suprathreshold psychoacoustical tuning curves at low and high frequencies. *J Assoc Res Otolaryngol*, 14(3), 341-357.

Backus BC, Guinan JJ (2006) Time-course of the human medial olivocochlear reflex. *J Acoust Soc Am*, 119(5), 2889-2904.

Boyd PJ (2006) Effects of programming threshold and maplaw settings on acoustic thresholds and speech discrimination with the MED-EL COMBI 40+ cochlear implant. *Ear Hear*, 27(6), 608-618.

Bregman AS (2001) Auditory Scene Analysis: The Perceptual Organization of Sound. Massachusets Institute of Technology: Second MIT Press.

Brown GJ, Ferry RT, Meddis R (2010) A computer model of auditory efferent suppression: implications for the recognition of speech in noise. *J Acoust Soc Am*, 127(2), 943-954.

Chintanpalli A, Jennings SG, Heinz MG, Strickland EA (2012) Modeling the anti-masking effects of the olivocochlear reflex in auditory nerve responses to tones in sustained noise. *J Assoc Res Otolaryngol*, 13(2), 219-235.

Clark NR, Brown GJ, Jurgens T, Meddis R (2012) A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise. *J Acoust Soc Am*, 132(3), 1535-1541.

Cooper NP, Guinan JJ (2006) Efferent-mediated control of basilar membrane motion. *J Physiol*, 576(Pt 1), 49-54.

Gardner WG, Martin KD (1995) HRTF measurements of a KEMAR. *J Acoust Soc Am*, 97(6), 3907-3908.

Guinan JJ (2006) Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear Hear*, 27(6), 589-607.

Guinan JJ (2010) Cochlear efferent innervation and function. *Curr Opin Otolaryngol Head Neck Surg*, 18(5), 447-453.

Hood LJ, Berlin CI, Hurley A, Cecola RP, Bell B (1996) Contralateral suppression of transient-evoked otoacoustic emissions in humans: intensity effects. *Hear Res*, 101(1-2), 113-118.

Kim SH, Frisina RD, Frisina DR (2006) Effects of age on speech understanding in normal hearing listeners: Relationship between the auditory efferent system and speech intelligibility in noise. *Speech Com*, 48(855), 862.

Lilaonitkul W, Guinan JJ (2009a) Human medial olivocochlear reflex: effects as functions of contralateral, ipsilateral, and bilateral elicitor bandwidths. *J Assoc Res Otolaryngol*, 10(3), 459-470.

Lilaonitkul W, Guinan JJ (2009b) Reflex control of the human inner ear: a half-octave offset in medial efferent feedback that is consistent with an efferent role in the control of masking. *J Neurophysiol*, 101(3), 1394-1406.

Maison SF, Usubuchi H, Liberman MC (2013) Efferent feedback minimizes cochlear neuropathy from moderate noise exposure. *J Neurosci*, 33(13), 5542-5552.

Robles L, Ruggero MA (2001) Mechanics of the mammalian cochlea. *Physiol Rev*, 81(3), 1305-1352.

Walsh EJ, McGee J, McFadden SL, Liberman MC (1998) Long-term effects of sectioning the olivocochlear bundle in neonatal cats. *J Neurosci*, 18(10), 3859-3869.

Wilson BS, Finley CC, Lawson DT, Wolford RD, Eddington DK, Rabinowitz WM (1991) Better speech recognition with cochlear implants. *Nature*, 352(6332), 236-238.

Chapter 13

Deactivating Cochlear Implant Electrodes Based on Pitch Information for Users of the ACE Strategy

Deborah Vickers, Aneeka Degun, Angela Canas
Ear Institute, University College London, London, UK

Thomas Stainsby, Filiep Vanpoucke
Cochlear Technology Centre, Mechelen, Belgium

Contact author: D. A. Vickers, d.vickers@ucl.ac.uk

Abstract

There is a wide range in performance for cochlear implant (CI) users and there is some evidence to suggest that implant fitting can be modified to improve performance if electrodes that do not provide distinct pitch information are de-activated. However, improvements in performance may not be the same for users of all CI devices; in particular for those with Cochlear devices using n-of-m strategies (ACE or SPEAK).

The goal of this research was to determine for users of Cochlear devices (CP810 implants with system 5 or 6 sound processors) if speech perception could be improved when indiscriminable electrodes were de-activated and this was also compared to when the same number of discriminable electrodes were de-activated.

A cross-over study was conducted with thirteen adult CI users who received experimental maps with de-activated channels for a minimum of two months and these were compared to an optimised clinical programme.

The findings showed that there were no significant benefits of electrode de-activation on speech perception and that there was a significant deterioration in spectro-temporal ripple perception when electrodes were switched off. There were no significant differences between de-activation of discriminable or indiscriminable electrodes.

These findings suggest that electrode de-activation with n-of-m strategies may not be beneficial.

Keywords: Cochlear implant, Electrode discrimination, Electrode deactivation, Fitting, Mapping, Electrode-Neurone Interface, Speech Perception, Pitch Ranking, ACE strategy, n-of-m

1 Introduction

Various approaches have been explored to look at the effects of re-mapping cochlear implants (CIs) to overcome the problems associated with neural dead regions or poor electrode placement (electrode-neural interface). The mapping interventions have typically explored ways to identify electrodes for de-activation using Computerized Tomography (CT), threshold profiling with focussed stimulation, detection of modulation and electrode discrimination based on pitch ranking. The electrodes indicated as being in a region with a poor electrode-neural

interface have typically been de-activated with the intention of improving spectral resolution by reducing overlapping electrode stimulation patterns.

The findings from electrode de-activation studies are mixed. Noble et al. (2014) used a CT-guided technique in which pre- and post-operative scans were used to determine the spatial positioning of the electrodes with respect to the modiolus. Poorly positioned electrodes were de-activated and additional mapping adjustments were made. Assessments were conducted at baseline and following a 3-6 week exposure period with the adjusted map. Learning effects were not accounted for in the design. Thirty six out of 72 (50%) participants demonstrated a benefit on one outcome measure, 17 had a significant degradation in performance (24%), 3 cochlear users had mixed results (4%) and the final 16 participants didn't demonstrate any effect of the de-activation (22%). When broken down by device the results showed that significant benefit on one measure was shown in 16 out of 26 AB and MEDEL users (62%; 7 showed significant deterioration) and 20 out of 46 Cochlear users (44%; 10 showed significant deterioration and 3 a mixture of significant deterioration and significant improvement). Overall the results demonstrated significant improvements in speech understanding in both quiet and noise for re-mapped participants, but the Cochlear device users were less likely to gain benefit from electrode de-activation than users of the other devices. It could be speculated that the n-of-m (n channels with highest input amplitudes are stimulated out of a possible m channels; e.g. ACE or SPEAK) strategies used in the Cochlear device may be less sensitive to channel overlap effects as the other device users because not all channels are stimulated in each cycle, thus making them less sensitive to channel overlap.

Zwolan et al. (1997) assessed users of an older generation Cochlear device (Nucleus mini-22) stimulated in bipolar mode using an MPEAK sound processing strategy. They demonstrated positive results when deactivating electrodes based on a pitch discrimination test. The testing was conducted acutely with care to randomise between clinical and experimental maps in the assessment, but the experimental map exposure was far briefer than that of the clinical map. They found significant improvements in speech perception in at least one speech test for 7/9 subjects; while 2/9 subjects showed a significant decrease on at least one speech test. This study demonstrated some benefit from channel de-activation for Cochlear users; however this was with an MPEAK strategy which was based on feature extraction and not an n-of-m approach.

Recent work by Saleh (Saleh et al., 2013; Saleh 2013) using listeners with more up-to-date sound processing strategies also demonstrated positive findings when de-activating electrodes. They based their channel selection approach on an acoustic pitch ranking technique. Electrodes were identified for de-activation if they did not provide distinct or tonotopic pitch information. Twenty five participants were enrolled. They received take-home experimental maps for a 1-month time period before being assessed. Assessments were conducted with the clinical map at baseline and again following exposure with the experimental map. They found statistically significant improvements for electrode de-activation on speech perception tests for 16 of the 25 participants (67%). For the Cochlear users 56% (5 out of 9) gained benefit compared to 69% for the AB and MEDEL users (11 out of 16). The numbers were small but the trend for fewer Cochlear users to receive benefit compared to other devices was similar to that observed by Noble et al., (2014).

Garadat et al (2013) demonstrated significant benefit with some users of the Cochlear device when electrodes were de-activated based on modulation detection. Up to five electrodes were switched off and they were selected such that only individual non-adjacent contacts were de-activated to avoiding large regions being switched off, drastically affecting filter allocation. Overall consonant and speech in noise perception improved, but vowel perception deteriorated. The testing was acute so learning effects were not controlled for and the listeners would not have adapted to the experimental map; with further experience and acclimatisation the users may have performed better on the vowel task, particular because it is known that vowel perception is sensitive to adjustments in filtering (Friesen et al., 1999). In general, these results

are positive for Cochlear users because the participants used the ACE strategy and demonstrated some degree of benefit from channel de-activation.

In contrast, some researchers have not observed benefit from electrode de-activation and have even demonstrated potentially deleterious effects when using the Cochlear device. Henshall and McKay (2001) used a multidimensional scaling technique to identify and de-activate electrodes with non-tonotopic percepts. They did not find an improvement in speech perception when channels were switched off and hypothesized that this could be due to the shift in frequency-to-electrode allocation produced by the clinical software when electrodes are deactivated; particularly in the apical region.

The findings from different electrode de-activation studies are inconclusive, partly due to the difficulties of comparing experimental maps to clinical maps that the CI users had adapted to and partly due to the impact that de-activating channels can have on rate of stimulation or filter-frequency allocation, which differ between devices and sound processing strategies. There is a trend suggesting that the impact of de-activating electrodes may be less effective for the Cochlear devices implementing ACE or SPEAK than it might be for the CIS based sound processing strategies incorporated in AB or MEDEL devices.

The goal of this research was to use pitch ranking tasks to identify indiscriminable electrodes for de-activation with users of latest generation Cochlear Devices and compare between the de-activation of indiscriminable and discriminable electrodes using a carefully controlled cross-over design. Discriminable electrodes were also de-activated to determine if any benefits observed were due to switching off electrodes with a poor electrode-neurone interface or just due to having fewer channels.

The research objectives were:

1. To determine if de-activating electrodes that are indiscriminable based on pitch ranking leads to changes in speech perception for adult CI users with the ACE strategy.
2. To determine if de-activating the same number of electrodes that were deemed discriminable based on pitch ranking leads to significantly different performance to the de-activation of indiscriminable electrodes.

2 Methods

2.1 Ethics Approval

Research was ethically approved by National Research Evaluation Service Committee North East (PRSC 23/07/2014).

2.2 Participants

Thirteen post-lingually deafened adults with Cochlear CP810 implants (using Nucleus 5 or 6 sound processors) using the ACE strategy were recruited. Individuals with ossification or fibrosis were excluded and participants had to have a minimum of 18 months CI experience and speak English as their primary language. The median age was 67 years; the median length of CI use was 2 years (mean 4 years). The median duration of severe-to-profound hearing loss prior to surgery was 4 years (mean 2 years).

2.3 Study design

A cross-over single-blinded (participant unaware of mapping interventions) randomised control trial was conducted. Prior to the crossover study each participant had fitting levels re-measured on a channel-by-channel basis to optimise their clinical map. They were given a minimum of one month to experience the new map prior to entering the crossover study. The crossover study

used an A-B-B-A / B-A-A-B design in which A was a map with indiscriminable electrodes de-activated and B was a map with discriminable electrodes de-activated.

2.4 Equipment

For the majority of the assessments the sound processor was placed inside an Otocube; a sound proof box designed for testing CI users to simulate listening over a loudspeaker in a sound treated room. All stimuli were delivered and the results recorded from a laptop.

The direct stimulation pitch-ranking task was conducted using the Nucleus Implant Communicator (NIC) to send stimuli directly to individual channels on the implant.

2.5 Test Materials

CHEAR Auditory Perception Test (CAPT)

The CAPT was used because it is sensitive to spectral differences in hearing aid fitting algorithms (Marriage et al., 2011) and the test re-test reliability was known (Vickers et al, 2013).

The CAPT is a four-alternative-forced-choice monosyllabic word discrimination test spoken by a female speaker of British English. It consisted of 10 sets of four minimally contrastive monosyllabic real words; eight sets with a contrastive consonant, e.g. fat, bat, cat, mat and two sets with a contrastive vowel e.g. cat, cot, cut, cart. The participant responded on a screen containing four pictures, each representing one of the words.

Stimuli were presented at 50dBA.

Children's Coordinate Response Measure (CCRM)

The adaptive CCRM sentence test was based on the test developed by Bolia et al. (2000) and Brungart (2001). It was used because of the relatively low contextual cues, and it was sufficiently straightforward for all participants to be assessed.

Each stimulus sentence took the form:

> 'Show the dog where the (colour) (number) is?': e.g. 'show the dog where the green three is?'

There were 6 colour options (blue, black, green, pink, red and white) and 8 possible numbers (1-9, excluding 7). The stimuli were spoken by a native British female speaker.

Two different noise competitors were used: speech-shaped babble or speech-shaped noise. The speech-shaped babble was based on male sentences. The amplitude-envelope of sentences was extracted by full-wave rectification and low-pass filtered at 30Hz. The envelope was used to modulate a speech-shaped noise.

The speech-shaped noise was a continuous section of random noise with the same long-term average spectrum as the speech materials. Random sections of the noise were selected.

The sentences were presented at 65dBA and the noise adjusted adaptively (2-down/1-up). Initial step size was 10 dB and final step size was 2dB and the final 6 reversals were averaged to obtain the speech reception threshold.

Spectral-Temporally Modulated Ripple Task (SMRT)

The spectral-temporally modulated ripple test (SMRT) was used to assess spectral resolution. The SMRT test was chosen over other spectral ripple tasks, because it avoided potential confounds such as loudness and edge-frequency cues whilst still being sensitive to spectral resolution (Aronoff and Landsberger, 2013).

Stimuli were 500 ms long with 100 ms ramps, nonharmonic tone complexes with 202 equal amplitude pure-tone frequency components, spaced every 1/33.3 octave from 100 to 6400 Hz. The amplitudes of the pure tones were modulated by a sine wave. A three-interval, two-alternative forced choice task was used, with a reference stimulus of 20 Ripples per octave (RPO) presented at 65 dBA. The target stimulus initially had 0.5 RPO and the number of ripples was modified using a 1-up/1-down adaptive procedure with a step size of 0.2 RPO. The test was completed after ten reversals. Thresholds were calculated based on the average of the last six reversals.

Pitch Ranking approaches

The determination of discriminable and indiscriminable electrodes was based on two approaches, the use of acoustic tones presented at the centre frequency of a filter via the System for Testing Auditory Responses (STAR) and direct stimulation using biphasic pulses to directly stimulate electrodes using the Nucleus Implant Communicator (NIC).

The pitch ranking task was the same for both delivery approaches. A two-interval two-alternative forced paradigm was used. A single presentation consisted of two intervals in which the higher tone was randomly assigned to either the first or second interval, and the listener indicated which interval sounded higher in pitch. The two tones/pulses were presented sequentially with a duration of 1000 ms and 500ms apart. Five pairs of tones/pulses were initially presented and responses recorded. If the participant got all five correct the pair was considered discriminable. If they scored less than 5 a further five presentations were given and scored. If the participant got 8 out of 10 presentations correct the pair passed and was considered discriminable. This was based on binomial significance at the $p< 0.05$ level (Skellam, 1948). A run consisted of presentation of all possible electrode pairs presented either five or ten times. For the STAR delivery the stimuli were presented at 65dBA and the NIC presentation was at 85% of the upper comfort level (C level). Level roving was applied to each presentation and this was +/- 3 dB for STAR and +/- 3 clinical units (CU) for NIC.

Results from the STAR and NIC testing were combined to identify indiscriminable electrodes. After determination of the "indiscriminable" electrode set, a "discriminable" electrode set of the same size was selected, using electrodes that were as near as possible to the indiscriminable set and had not failed on either task.

Fitting procedure for de-activation

Each electrode was switched off in turn, until the entire A/B set was de-activated. Overall level was checked to ensure it was perceived to be equally loud to the optimised clinical map.

3 Results

Non-parametric statics were used due to the small number of participants (N=13) and the data for the CCRM (babble) and the SMRT exceeded the acceptable limits for skewness and kurtosis. Each mapping condition was tested on two occasions and the highest score was used in the analysis.

Figure 1 shows the group results for all four of the outcome measures, comparing the optimised clinical map to the "indiscriminable" and "discriminable" de-activation conditions.

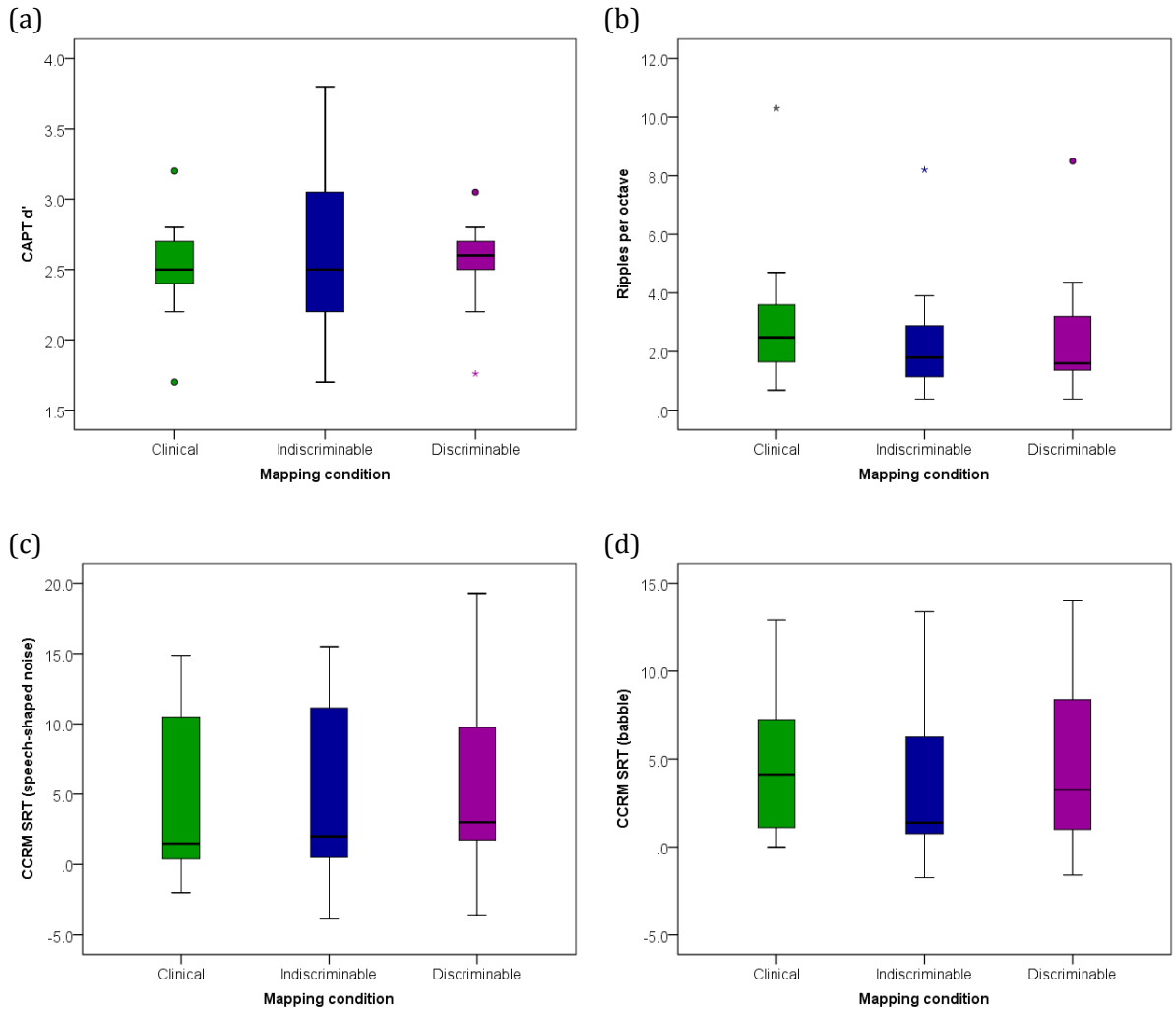


Figure 1. Boxplots for group results for CAPT d' (panel a), SMRT ripples per octave (panel b), CCRM in speech-shaped noise (panel c), CCRM in babble (panel d). The boxes represents the inter-quartile range, the line in the box shows the median, and the whiskers indicate the range of results and the asterisks and circles show any outliers.

For the CAPT d', CCRM (speech-shaped), CCRM (babble), scores there were no significant differences between any of the mapping options at the $p < 0.05$ level. The only significant differences that arose were for the SMRT which was significantly poorer for the indiscriminable electrodes de-activated than the clinical map ($z=-2.9$, $p=0.004$) and was also significantly poorer when the discriminable electrodes were de-activated ($z=-2.62$, $p=0.009$), there was no significant difference between the de-activation of the indiscriminable or discriminable electrodes.

4 Discussion

Results indicated that there was no measurable benefit for de-activating channels based on pitch ranking for Cochlear devices users with the ACE strategy. The effects were negligible regardless of whether the de-activated channels were indiscriminable or discriminable based on pitch ranking. Great care was taken to balance the design and ensure that all maps were used for sufficient time to obtain benefit. Prior to experimental map adjustments clinical maps were optimised and each participant was given time to adjust to this prior to starting the study.

Many participants (8 individuals) already had at least one electrode de-activated in their clinical map.

There was a range in the number of channels selected for de-activation in the experimental maps (between 1 to 6; median = four channels), which resulted in an overall range in de-activated channels of 1 to 10 (median = six channels). For an n-of-m strategy this number of de-activated electrodes may have been too small to demonstrate differences in performance; the selection of channels to de-activate was based purely on electrodes that were indiscriminable from one another in pitch ranking and for many users the number of indiscriminable electrode pairs was low so only small numbers were used in the de-activated maps.

The only statistically significant finding indicated that spectral resolution was poorer when electrodes were de-activated. When the channels were de-activated the rate of stimulation never changed, however the default filter allocation was used, which may have affected the spectral information (Henshall and McKay, 2001). For the de-activation of discriminable channels, the selected channels were as near in site to the de-activated indiscriminable channels in an attempt to avoid the filter allocations be dramatically different between the two conditions. However, for many people the changes to the frequency-to-electrode allocation would have been large.

There should be further exploration of the fitting parameter space for users of n-of-m strategies. The limits or extent of effective use for channel de-activation should be determined and predictive factors to aid the fitting process defined. It is also essential to determine the most critical factors to modify for those with poorer performance levels.

Acknowledgments

This project was funded by Cochlear Corporation.

References

Aronoff, J. M., & Landsberger, D. M. (2013). The development of a modified spectral ripple test. *Journal of Acoustical Society of America*, 134(2), EL217–EL222.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of Acoustical Society of America* 109(3):1101-1109.

Bolia, R. S., Nelson, W. T., Ericson, M. A. & Simpson B. D. (2000). A speech corpus for multitalker communications research. *Journal of Acoustical Society of America* 107(2):1065-1066

Friesen, L. M., Shannon, R. V. & Slattery, W. H. (1999). The effect of frequency allocation on phoneme recognition with the Nucleus-22 cochlear implant. *American Journal of Otology* 20:729-734

Garadat, S. N., Zwolan, T. A. & Pfingst, B. E. (2013). Using temporal modulation sensitivity to select stimulation sites for processor MAPs in cochlear implant listeners. *Audiology & Neuro-Otology* 18(4):247-260.

Henshall, K. R. & McKay, C. M. (2001). Optimizing electrode and filter selection in cochlear implant speech processor maps. *Journal American Academy of Audiology*. 12(9):478-489.

Marriage, J. E., Vickers, D. A., Baer, T., & Moore, B. C. J. (2011). Using speech perception measures to guide the choice of amplification. In R. C. Seewald & J. M. Bamford (Eds.). *A Sound Foundation through Early Amplification.* Staefa: Phonak. 273-279.

Noble, J. H., Gifford, R.H., Hedley-Williams A. J., Dawant, B. M., Labadie, R. F. (2014). Clinical evaluation of an image-guided cochlear implant programming strategy. *Audiology & Neuro-Otology* 19(6):400-411.

Saleh, S. M. I. (2013). Fitting Cochlear Implants Based on Pitch Perception. *Doctoral Dissertation*. University College London.

Saleh, S. M., Saeed, S. R., Meerton, L., Moore, D. R., & Vickers, D. A. (2013). Clinical use of electrode differentiation to enhance programming of cochlear implants. Cochlear Implants International 14: S16–S18.

Skellam, J. G. (1948). A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success As Variable Between the Sets of Trials. Journal of the Royal Statistical Society Series B-Statistical Methodology, 10(2): 257-261.

Vickers, D., Backus, B., Macdonald, N., Rostamzadeh, N., Marriage, J., and Mahon, M. (2013) Using personal response systems to assess speech perception within the classroom: an approach to determine the efficacy of sound field re-enforcement in primary school classrooms. *Ear & Hearing* 34:491–502

Zwolan, T. A., Collins, L. M., & Wakefield, G. H. (1997). Electrode discrimination and speech recognition in postlingually deafened adult cochlear implant subjects. *Journal Acoustical Society of America* 102(6): 3673-3685.

Chapter 14

Speech Masking and Masking Release in Normal, Impaired and Electric Hearing: Interactions Between Frequency Selectivity and Inherent Temporal Fluctuations in Noise

Andrew J. Oxenham
Departments of Psychology and Otolaryngology, University of Minnesota – Twin Cities, Minneapolis, MN, USA

Heather A. Kreft
Department of Otolaryngology, University of Minnesota – Twin Cities, Minneapolis, MN, USA

Contact author: AJO, oxenham@umn.edu

Abstract

Recent studies in normal-hearing listeners have used envelope-vocoded stimuli to show that the masking of speech by noise is dominated by the temporal-envelope fluctuations inherent in noise, rather than just the overall power of the noise. Because these studies were based on vocoding, it was expected that cochlear-implant (CI) users would demonstrate a similar sensitivity to inherent fluctuations. In contrast, we found that CI users showed no difference in speech intelligibility between maskers with inherent envelope fluctuations, and maskers with no fluctuations. Here, these initial findings in CI users were extended to listeners with cochlear hearing loss and the results were compared with those from normal-hearing listeners at either equal sensation level or equal sound pressure level. In CI users, the results support the predictions based of the hypothesis that poor spectral resolution leads to a summation of independently fluctuating temporal envelopes, which in turn leads to "smoother" effective noise temporal envelopes with less energy at modulation frequencies relevant for speech. In hearing-impaired listeners (and in normal-hearing listeners at high sound levels), the results are consistent with a reduction in low-frequency inherent noise fluctuations due to broader cochlear filtering. The reduced effect of inherent temporal fluctuations in noise, due to either current spread or broader cochlear filters, provides a new way to explain the loss of masking release experienced in CI users and hearing-impaired listeners when additional amplitude fluctuations are introduced in noise maskers.

Keywords: cochlear hearing loss, hearing in noise, speech perception.

1 Introduction

Speech perception is a major communication challenge for people with hearing loss and with cochlear implants (CIs), particularly when the speech is embedded in background noise (e.g., Humes et al., 2002; Zeng, 2004). Noise reduction algorithms are sometimes used in both hearing

aids and CIs but, although they may improve the physical signal-to-noise ratio, they rarely produce improvements in speech intelligibility (e.g., Jorgensen et al., 2013).

Recent work has suggested that it is not so much the overall noise energy that limits speech perception in noise, as suggested by earlier work (French & Steinberg, 1947; Kryter, 1962; George et al., 2008), but rather the energy in the inherent temporal-envelope modulation in noise (Dubbelboer & Houtgast, 2008; Jorgensen & Dau, 2011; Stone et al., 2011; Stone et al., 2012; Jorgensen et al., 2013; Stone & Moore, 2014). This shift in emphasis from energy to modulation energy suggests that noise-reduction algorithms might do better if they aim at reducing the temporal fluctuations in noise, rather than simply reducing the overall noise energy. However, the work mentioned so far on the role of inherent temporal-envelope fluctuations was carried out in normal-hearing (NH) listeners, using tone-vocoder techniques to simulate certain aspects of CI processing (Shannon et al., 1995), rather than in the target populations of hearing-impaired (HI) listeners and CI users.

In a recent study (Oxenham & Kreft, 2014) we examined the effects of inherent noise fluctuations in CI users. As CI users are reliant primarily on temporal-envelope cues, we expected to find the same pattern of results that had been reported in NH listeners, with masking being dominated by the inherent noise fluctuations, and maskers without inherent fluctuations producing much less masking. Instead, we found that CI users exhibited no benefit of maskers without inherent fluctuations. In other words, the assumption that masker modulation energy determined speech perception in noise did not hold for CI users. After ruling out the possibility that CI users were insensitive to amplitude modulation, we tested the new hypothesis that the effective inherent noise envelope fluctuations were reduced in the CI users, due to the effects of current spread, or interaction between adjacent electrodes. The idea is that the independent envelope fluctuations produced by the noise at the output of each CI channel interact with the fluctuations in adjacent channels to produce an envelope that is smoother overall. This hypothesis was tested by simulating current spread using the tone vocoders in NH listeners. With simulated current spread, we found very similar results in NH listeners to the ones we obtained in CI users, including the loss of any effect of inherent noise fluctuations.

One remaining question is whether HI listeners exhibit the same loss of sensitivity to inherent masker fluctuations as CI users. If so, the finding may go some way to explaining why both HI and CI populations exhibit less masking release than NH listeners when additional slow fluctuations are imposed on noise maskers (e.g., Festen & Plomp, 1990; Nelson & Jin, 2004; Stickney et al., 2004; Gregan et al., 2013). On one hand, cochlear hearing loss is generally accompanied by a loss of frequency selectivity, due to loss of function of the outer hair cells; on the other hand this "smearing" of the spectrum occurs before extraction of the temporal envelope, rather than afterwards, as is the case with CI processing. Because the smearing is due to wider filters, rather than envelope summation, the resultant envelopes from broader filters still have Rayleigh-distributed envelopes with the same overall relative modulation power as the envelopes from narrower filters. In contrast, with CIs, the envelopes derived from summing the envelope currents from adjacent electrodes results in envelopes that are no longer Rayleigh distributed and have lower modulation power, relative to the overall (DC) power in the envelope.

One factor suggesting that HI listeners may also experience less influence of inherent noise fluctuations is that the modulation spectrum is altered by broadening the filters: for an ideal rectangular filter, the modulation power of Gaussian noise after filtering has a triangular distribution, with a maximum at 0 Hz and reaching a minimum of no power at a frequency equal to the bandwidth of the filter (Lawson & Uhlenbeck, 1950). Although widening the filter does not alter the area under the modulation spectrum, it results in relatively less power at lower modulation frequencies because the modulation spectrum now extends out to higher modulation frequencies, thereby "flattening" the triangle (see Fig. 1). Given that the low modulation frequencies are most important for speech, the relative reduction in modulation

power at low modulation frequencies may reduce the influence of the inherent fluctuations for listeners with broader filters, due to hearing loss.

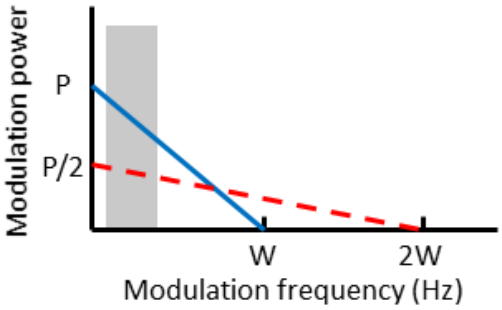


Figure 1. Schematic diagram of the effect of broadening the filter from a bandwidth of W to a bandwidth of $2W$ on the modulation spectrum of filtered Gaussian noise. The relative modulation power (area under the line) remains constant, but the area under the lines within the speech-relevant range (shaded rectangle) is reduced.

The aim of this experiment was to test the hypothesis that hearing loss leads to less effect of inherent noise fluctuations on speech masking. We compared the results of listeners with cochlear hearing loss with the performance of young NH listeners, and age-matched NH listeners. Performance was compared for roughly equal sensation levels (SL), and for equal sound pressure levels (SPL) to test for the effect of overall level on performance in NH listeners.

2 Methods

2.1 Listeners

Nine listeners with mild-to-moderate sensorineural hearing loss (4 male and 5 female; mean age 61.2 years) took part in this experiment. Their 4-frequency pure-tone average thresholds (4F-PTA from 500, 1000, 2000, and 4000 Hz) are shown in Figure 2. Nine listeners with clinically normal hearing (defined as 20 dB HL or less at octave frequencies between 250 Hz and 4000 Hz; mean 4F-PTA 7.6 dB HL), who were matched for age (mean age 62.2 years) and gender with the HI listeners, were run as the primary comparison group. In addition, a group of four young (mean age 20.5 years; mean 4F-PTA 2.8 dB HL) NH listeners were tested. All experimental protocols were approved by the Institutional Review Board of the University of Minnesota, and all listeners provided informed written consent prior to participation.

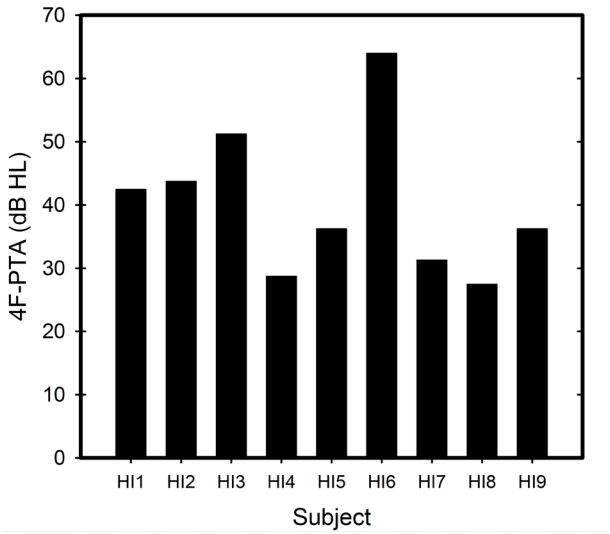


Figure 2. 4F-PTA (in dB HL). The 4F-PTA calculated by averaging the 500, 1000, 2000, and 4000 Hz audiometric thresholds for each HI subject.

2.2 Stimuli

Listeners were presented with sentences taken from the AzBio speech corpus (Spahr et al., 2012). The sentences were presented to the HI listeners at an overall rms level of 85 dB SPL. The sentences were presented to the NH listeners at two different levels: in the equal-SL condition, the sentences were presented at 40 dB above the detection threshold of the speech for the NH listeners, similar to the sensation level (SL) of the 85-dB SPL speech for the HI listeners; in the equal-SPL condition, the speech was presented at 85 dB SPL. The sentences were presented in three different types of masker, as in Oxenham and Kreft (2014); see Figure 3. The first was a Gaussian noise, spectrally shaped to match the long-term spectrum of the target speech. The second comprised of 16 pure tones, approximately evenly spaced on a logarithmic frequency scale from 333 to 6665 Hz, corresponding to the center frequencies on a standard CI map for Advanced Bionics. The amplitudes of the tones were selected to produce the same long-term output of a 16-channel vocoder as the target speech (or Gaussian noise masker), assuming the same center frequencies. The third was comprised of the same 16 tones, but each tone was modulated independently with the temporal envelope of a noise masker, bandpass filtered using a vocoder filter with the same center frequency as the tone carrier. These three maskers produced equal amounts of speech masking in the CI users tested by Oxenham and Kreft (2014). The masker was gated on 1 s before the beginning of each sentence, and was gated off 1 s after the end of each sentence. The masker in each trial was a sample of a longer 25-s sound file, cut randomly from within that longer waveform. The speech and masker were mixed before presentation and the signal-to-masker ratios were selected in advance, based on pilot data, to span a range of performance between 0 and 100% word recognition.

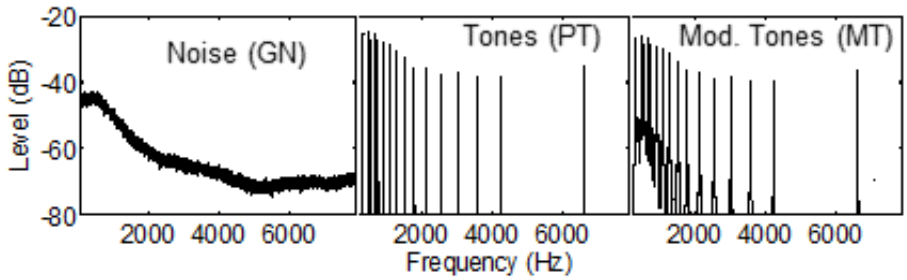


Figure 3. Representation of the three masker types used in the experiment. The three panels provide spectral representations of the noise (top), tone (middle), and modulated tone (bottom) maskers.

The speech and the masker were mixed and low-pass filtered at 4000-Hz, and were either presented unprocessed or were passed through a tone-excited envelope vocoder that simulates certain aspects of CI processing (Dorman et al., 1998; Whitmal et al., 2007). The stimulus was divided into 16 frequency subbands, with the same center frequencies as the 16 tone maskers. The temporal envelope from each subband was extracted using a Hilbert transform, and then the resulting envelope was lowpass filtered with a 4th-order Butterworth filter and a cutoff frequency of 50 Hz. This cutoff frequency was chosen to reduce possible voicing periodicity cues, and to reduce the possibility that the vocoding produced spectrally resolved components via the amplitude modulation. Each temporal envelope was then used to modulate a pure tone at the center frequency of the respective subband.

2.3 Procedure

The stimuli were generated and processed using Matlab (Mathworks, Natick, MA). The sounds were converted via a 24-bit digital-to-analog converter and were presented via headphones. The stimuli were presented to one ear (the better each in the HI listeners), and the speech-shaped noise was presented in the opposite ear at a level 30 dB below the level of the speech. The

listeners were seated individually in a double-walled sound-attenuating booth. Listeners responded to sentences by typing what they heard via a computer keyboard. They were encouraged to guess individual words, even if they had not heard or understood the entire sentence. Sentences were scored for words correct as a proportion of the total number of keywords presented. One sentence list (of 20 sentences) was completed for each masker type and masker level. Presentation was blocked by condition (natural and vocoded, and speech level), and the order of presentation was counterbalanced across listeners. The test order of signal-to-masker ratio was random within each block.

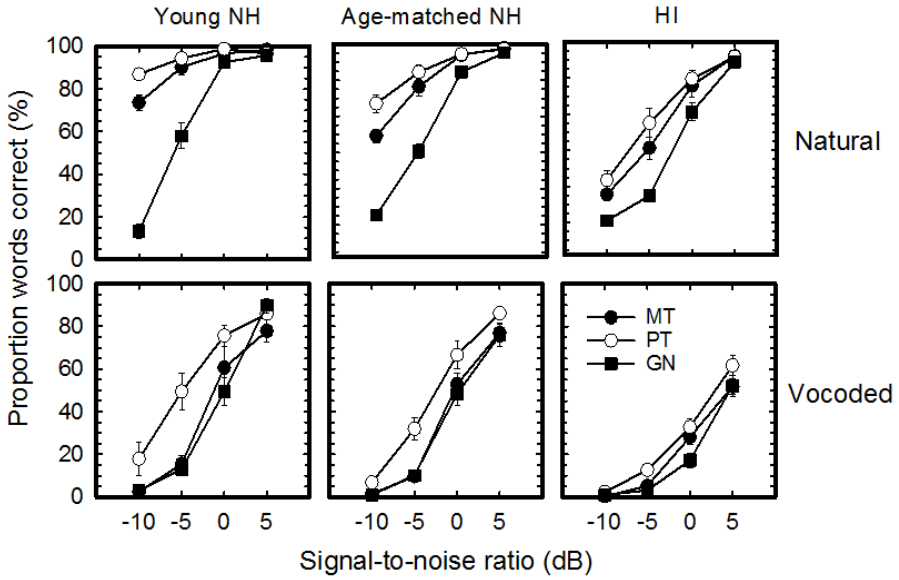


Figure 4. Speech intelligibility with the speech at roughly equal sensation level (SL) across the three groups. Sentence recognition was tested in speech-shaped Gaussian noise (GN), pure tones (PT), and modulated tones (MT).

3 Results

The results from the conditions where the speech was presented at roughly equal sensation level to all participants (85 dB SPL for the HI group, and 40 dB SL for the NH groups) are shown in Figure 4. The upper row shows results without vocoding; the lower row shows results with vocoding. The results from the young NH, age-matched NH, and HI listeners are shown in the left, middle and right columns, respectively.

Consider first the results from the non-vocoded conditions (upper row of Fig. 4). Performance in the presence of Gaussian noise (GN; filled squares) is quite similar across the three groups but, as expected, performance is slightly poorer in the HI group. When the noise masker is replaced by 16 pure tones (PT; open circles), performance is markedly better – all three groups show a large improvement in performance, although there appears to be a systematic decrease in performance from young NH to age-matched NH to the HI group. Imposing noise envelope fluctuations to the pure-tone maskers (MT; filled circles) results in a smaller reduction in performance for all three groups. It appears that both the spectral sparsity of the tonal maskers (PT and MT), as well as the lack of fluctuations (PT) contributes to the release from masking, relative to the GN condition.

Consider next the results from the vocoded conditions (lower row of Fig. 4). Here, most benefits of frequency selectivity have been reduced by limiting spectral resolution to the 16 vocoder channels. As expected, the MT masker produces very similar results to the GN masker, as both produce very similar outputs from the tone vocoder. The young NH and the age-matched NH

groups seem able to take similar advantage of the lack of inherent masker fluctuations in the PT condition. In contrast, the differences between the PT and MT seem less pronounced in the HI listeners; although some differences remain (in contrast to CI users), they are smaller than in the NH listeners.

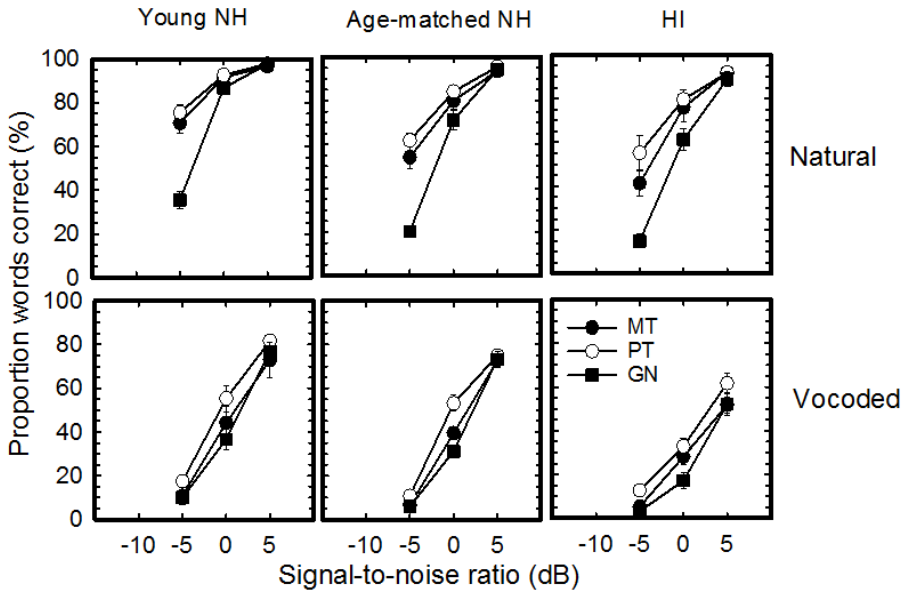


Figure 5. Speech intelligibility with the speech at equal sound pressure level (85 dB SPL). Data from the HI listeners are replotted from Fig. 4.

The results using 85-dB SPL speech for all listeners are shown in Figure 5. Only three speech-to-masker ratios (-5, 0, and 5 dB) were tested in this part of the experiment. The relevant data from the HI listeners are simply replotted from Figure 4. Performance was generally worse overall for both the young and age-matched NH listeners, compared to results at the lower sound level. In addition, the difference between the PT and MT conditions was reduced. This reduction was observed even in the non-vocoded conditions, but was particularly apparent in the vocoded conditions. In fact, in the vocoded conditions, the differences between the NH groups and the HI group were relatively small.

4 Discussion

This experiment was designed to test whether the loss of frequency selectivity due to cochlear hearing loss also resulted in a loss of sensitivity to the effects of inherent masker fluctuations in speech understanding. The results from HI listeners were compared with those from both young NH listeners and age-matched NH listeners. Overall, the HI group showed smaller-than-normal differences between maskers with and without inherent fluctuations. This loss of sensitivity to inherent fluctuations was particularly apparent in the vocoded conditions (Fig. 4; lower right panel). However, in contrast to earlier results from CI users (Oxenham & Kreft, 2014), some differences remained between conditions with and without inherent fluctuations, suggesting that the effects of poorer frequency selectivity are not as profound as for CI users. This could be for two reasons: First, frequency selectivity in HI listeners with mild-to-moderate hearing loss is not as poor as that in CI users, where there is generally very strong interaction between adjacent (and non-adjacent) electrodes). Second, the difference between the interaction occurring before and after envelope extraction may affect outcomes to some extent, in that the reduction in masker modulation power may be greater for CI users than for HI listeners, even with a similar loss of frequency selectivity.

One interesting outcome was that the differences between the NH groups and the HI group were not very pronounced when the speech was presented to the groups at the same high SPL. It is well known that frequency selectivity becomes poorer in NH listeners at high levels (e.g., Nelson & Freyman, 1984; Glasberg & Moore, 2000). It may be that the broadening with level associated with NH may be sufficient to reduce the effects of inherent masker fluctuations.

Overall, the results show that the importance of inherent masker fluctuations in determined speech intelligibility in noise depends to some extent on the conditions and the listening population. It cannot be claimed that inherent masker fluctuations always limit speech perception, as the effect of the fluctuations is non-existence in CI users, and is greatly reduced in HI listeners and in NH listeners at high sound levels.

Acknowledgments

This work was supported by NIH grant R01 DC012262.

References

Dorman, M. F., Loizou, P. C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels. *Journal of the Acoustical Society of America, 104*(6), 3583-3585.

Dubbelboer, F., & Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *Journal of the Acoustical Society of America, 124*(6), 3937-3946. doi: 10.1121/1.3001713

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America, 88*(4), 1725-1736.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19*, 90-119.

George, E. L., Festen, J. M., & Houtgast, T. (2008). The combined effects of reverberation and nonstationary noise on sentence intelligibility. *Journal of the Acoustical Society of America, 124*(2), 1269-1277. doi: 10.1121/1.2945153

Glasberg, B. R., & Moore, B. C. J. (2000). Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *Journal of the Acoustical Society of America, 108*, 2318-2328.

Gregan, M. J., Nelson, P. B., & Oxenham, A. J. (2013). Behavioral measures of cochlear compression and temporal resolution as predictors of speech masking release in hearing-impaired listeners. *Journal of the Acoustical Society of America, 134*(4), 2895-2912. doi: 10.1121/1.4818773

Humes, L. E., Wilson, D. L., Barlow, N. N., & Garner, C. (2002). Changes in hearing-aid benefit following 1 or 2 years of hearing-aid use by older adults. *Journal of Speech, Language, and Hearing Research, 45*(4), 772-782.

Jorgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *Journal of the Acoustical Society of America, 130*(3), 1475-1487. doi: 10.1121/1.3621502

Jorgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America, 134*(1), 436-446. doi: 10.1121/1.4807563

Kryter, K. D. (1962). Methods for the calculation and use of the Articulation Index. *Journal of the Acoustical Society of America, 34*, 467-477.

Lawson, J. L., & Uhlenbeck, G. E. (1950). *Threshold Signals* (Vol. 24). New York: McGraw Hill.

Nelson, D. A., & Freyman, R. L. (1984). Broadened forward-masked tuning curves from intense masking tones: delay-time and probe level manipulations. *Journal of the Acoustical Society of America, 75*, 1570-1577.

Nelson, P. B., & Jin, S. H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America, 115*(5 Pt 1), 2286-2294.

Oxenham, A. J., & Kreft, H. A. (2014). Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing. *Trends Hear, 18*. doi: 10.1177/2331216514553783

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., & Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing, 33*(1), 112-117. doi: 10.1097/AUD.0b013e31822c2549

Stickney, G. S., Zeng, F. G., Litovsky, R., & Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *Journal of the Acoustical Society of America, 116*(2), 1081-1091.

Stone, M. A., & Moore, B. C. J. (2014). On the near non-existence of "pure" energetic masking release for speech. *Journal of the Acoustical Society of America, 135*(4), 1967-1977.

Stone, M. A., Fullgrabe, C., & Moore, B. C. J. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *Journal of the Acoustical Society of America, 132*(1), 317-326. doi: 10.1121/1.4725766

Stone, M. A., Fullgrabe, C., Mackinnon, R. C., & Moore, B. C. J. (2011). The importance for speech intelligibility of random fluctuations in "steady" background noise. *Journal of the Acoustical Society of America, 130*(5), 2874-2881. doi: 10.1121/1.3641371

Whitmal, N. A., Poissant, S. F., Freyman, R. L., & Helfer, K. S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *Journal of the Acoustical Society of America, 122*, 2376-2388.

Zeng, F. G. (2004). Trends in cochlear implants. *Trends Amplif, 8*(1), 1-34.

Chapter 15

Effects of Pulse Shape and Polarity on Sensitivity to Cochlear Implant Stimulation: a Chronic Study in Guinea Pigs

Olivier Macherey
LMA-CNRS, UPR 7051, Aix-Marseille University, Centrale Marseille, Marseille, France

Yves Cazals
LNIA-CNRS, UMR 7260, Aix-Marseille University, Marseille, France

Contact author: OM, macherey@lma.cnrs-mrs.fr

Abstract

Most cochlear implants (CIs) stimulate the auditory nerve with trains of symmetric biphasic pulses consisting of two phases of opposite polarity. Animal and human studies have shown that both polarities can elicit neural responses. In human CI listeners, studies have shown that at suprathreshold levels, the anodic phase is more effective than the cathodic phase. In contrast, animal studies usually show the opposite trend. Although the reason for this discrepancy still remains unclear, computational modelling results have proposed that the degeneration of the peripheral processes of the neurons could lead to a higher efficiency of anodic stimulation. We tested this hypothesis in ten guinea pigs who were deafened with an injection of sysomycin and implanted with a single ball electrode inserted in the first turn of the cochlea. Animals were tested at regular intervals between 1 week after deafening and up to 1 year for some of them. Our hypothesis was that if the effect of polarity is determined by the presence or absence of peripheral processes, the difference in polarity efficiency should change over time because of a progressive neural degeneration. Stimuli consisted of charge-balanced symmetric and asymmetric pulses allowing us to observe the response to each polarity individually. For all stimuli, the inferior colliculus evoked potential was measured. Results show that the cathodic phase was more effective than the anodic phase and that this remained so even several months after deafening. This suggests that neural degeneration cannot entirely account for the higher efficiency of anodic stimulation observed in human CI listeners.

Keywords: inferior colliculus, electrical stimulation, asymmetric pulses, auditory nerve

1 Introduction

Despite the 40-years existence of cochlear implants (CIs), the basic mechanisms of auditory nerve electrical stimulation still remain partly unclear. Animal electrophysiological recordings may help better understand these mechanisms and guide the development of novel coding strategies while, at the same time, providing important data to validate computational models of the implanted cochlea. However, it appears necessary to evaluate the relevance of animal models with regards to psychophysical data measured in human CI listeners. Specifically

understanding the discrepancies between species appears as a pre-requesite to the development of more valuable animal models.

Most contemporary CIs stimulate the auditory nerve with trains of symmetric biphasic pulses consisting of two phases of opposite polarity presented in short succession. Although both polarities can elicit neural responses (Hartmann et al., 1984; Miller et al., 1998, 1999; Klop et al., 2004; Undurraga et al., 2013), they are not equally efficient. Animal recordings performed at the level of the auditory nerve usually show longer latencies and lower thresholds for cathodic than for anodic stimulation (e.g., Miller et al., 1999). This observation is consistent with theoretical models of extracellular electrical stimulation of the nerve showing that a cathodic current produces a larger depolarization and elicits action potentials more peripherally than an anodic current of the same level (Rattay, 1989).

In past years, studies performed with human CI listeners have shown different results. While at hearing threshold, no consistent difference between polarities has been observed, a large number of studies have reported that anodic stimulation is more effective than cathodic stimulation at suprathreshold levels (Macherey et al., 2006, 2008, 2011; Undurraga et al., 2010, 2013; Carlyon et al., 2013). This has been shown psychophysically in loudness balancing, pitch ranking and masking experiments and electrophysiologically in measures of electrically-evoked compound action potentials and auditory brainstem responses. Although this result contrasts with the conclusions of most animal reports, an exception is the study by Miller et al. (1998) who found a similar trend in their guinea pigs. Because the results of this specific study are at odds with other measurements performed in the guinea pig (Klop et al., 2004) and with all measurements performed in cats, the first aim of the present study was to clarify the effect of polarity in guinea pigs. This was done by measuring the inferior colliculus evoked potential in response to intracochlear electrical stimulation using various pulse shapes differing in polarity and in current level.

While the reason for the higher sensitivity of human CI users to anodic stimulation remains unknown, predictions from two different -among the few- computational models of the implanted cochlea offer potential explanations (Macherey et al., 2006). First, the modeling results of Rattay et al. (2001) suggest that polarity sensitivity may strongly depend on the state of degeneration and demyelination of the peripheral processes of the neurons. Specifically, the model predicts lower anodic thresholds for degenerated peripheral processes. Although post-mortem studies have shown that such degeneration can occur in CI listeners (Fayad & Linthicum, 2006), there is no currently available method to evaluate the presence of peripheral processes in patients. The second aim of this study was to study polarity sensitivity in guinea pigs as a function of time after deafening. Our reasoning was based on the observation that spiral ganglion neurons degenerate over time in deafened animals and that this degeneration first affects the peripheral processes (Koitchev et al., 1982, Cazals et al., 1983, Leake & Hradek, 1988). If this degeneration is responsible for the higher anodic sensitivity, we hypothesized that the effect of polarity would change over time.

An alternative explanation for higher anodic sensitivity at high levels comes from a model of the implanted guinea-pig cochlea (Frijns et al., 1996). This model predicts that cathodic stimulation should elicit action potentials at the periphery of the neurons but that above a certain current level, the propagation of these action potentials may be blocked by the hyperpolarization produced more centrally on the fibers. This blocking was not observed in simulations using anodic stimuli because action potentials were in this case elicited at a more central locus on the fibers. A third aim of the present study was to investigate if this behavior could be observed at the level of the inferior colliculus of the guinea pigs. If, at some level in the dynamic range, action potentials are being blocked for cathodic stimulation, this should be reflected by a slower rate of increase of the neural response for cathodic than for anodic stimulation and/or potentially by a non-monotonic growth of response amplitude as a function of current level.

Finally, most animal studies on polarity sensitivity have been performed using monophasic pulses which cannot safely be used in humans. This is why studies in CI subjects use charge-balanced asymmetric waveforms which are hypothesized to make one polarity dominate over the other. Some of these waveforms have the advantage of requiring lower current levels than the commonly-used symmetric waveform to elicit auditory sensations in implant listeners (Macherey et al., 2006). The fourth aim of the present study was to evaluate the efficiency of these different pulse shapes and to compare it to human psychophysical data obtained previously.

2 Methods

2.1 Subjects

Ten Albino guinea pigs (G1-G10) of the Dunkin-Hartley strain (obtained from the Charles River Company) were used in these studies. They weighed about 300 grams at the beginning of the experiment. They were kept in usual animal quarters with food and drink ad libitum and under a 12-hours darkness / 12-hours light cycle. The experimental procedures were carried out in accordance with French national legislation (JO 87-848, European Communities Council Directive (2010/63/EU,74) and local ethics committee “Direction Départementale de la Protection des Populations", with permit number 00002.01.

2.2 Animal preparation

In a first step, animals were chronically implanted with an electrode in the right inferior colliculus (IC). This technique allows easy and stable recordings over months as indicated in several previous experiments (Popelar et al., 1994, Arnold & Burkard, 1998, Deng et al., 2010) while microelectrodes recordings are much more cumbersome and cannot yield recordings from the same cells from one session to the next. The animal was anesthetized with a mixture of ketamine at 50 mg/kg and xylazine at 10 mg/kg injected in the thigh muscle. The head skull was incised so that the bregma and lambda sutures as well as the parietal bones were seen. They were then gently scraped and three small screws were placed in front of the bregma suture into the frontal bones which served to securely fix the head to an orientable holder. The skull was then adjusted so that the bregma and lambda suture points, and also the left and right sides were on a horizontal line so as to use stereotaxic coordinates. At the antero-posterior and lateral stereotaxic coordinates of the IC, a small hole was pierced through the skull. Using a micromanipulator, a Teflon-coated platinum electrode (127 micrometer in diameter obtained from A-M systems) was progressively lowered into the brain. The bare section at the tip of this wire served for recording. During the lowering into the brain, an acoustically-evoked potential was recorded using a click stimulus (0.1-ms rectangular pulse) delivered with a small headphone close to the animal’s left ear. The electrode depth into the brain was adjusted to the place which produced a maximal amplitude of IC evoked potential. Then the electrode was fixed to the skull with acrylic cement and soldered to a small connector to which two screws in the frontal bones were also connected and served as reference and ground, respectively. The ensemble was finally fixed to the surface of the skull with acrylic cement. In three animals (G8-G10), several electrodes were simultaneously implanted at different depths to examine possible variations in evoked responses at different levels of the IC.

After a recovery period of about one week, the animal was sedated with xylazine (5mg/kg), a small loudspeaker (MF1 from TDT) was placed at 1 cm from its left ear pinna and the animal’s head connector was linked to the input of an amplifier (A-M systems model 1700). Tone pips with 2 milliseconds linear rise/fall times at octave frequencies between 2 to 32 kHz were presented at levels of 90 to 0 dB SPL in 10-dB steps. In response to sound stimuli, electrophysiological responses were amplified 10000 times and filtered between 100 Hz and 5

kHz, then they were averaged over 100 to 300 times (CED 1401 A/D converter). All averaged responses were stored in hard disk memory for further use.

About a week later, the animal was anesthetized and the left ear bulla was approached from behind the pinna. A small hole was pierced in the bulla which allowed seeing the round window. The round window was gently pierced and a mixture of sisomycin at a concentration of 50 mg/ml was injected into the round window niche until the bulla was completely filled with the mixture. The animal was left so for two hours and the sisomycin was then aspirated from the bulla. This sisomycin treatment is known from previous experiments (Cazals et al., 1983) to completely destroy cochlear hair cells and lead to progressive spiral ganglion cells degeneration over months. Then a teflon-coated platinum wire (200 micrometers in diameter) with a tip bare over about 0.5 millimeter was introduced through the round window over a length of about 2 millimeters, so that the bare tip of this cochlear implant electrode was approximately situated at the tonotopic place corresponding to frequencies of 16-32 kHz. Because of deficiencies in handling and care in animals' quarter during the months of the experiment, for some animals, the head connector came off the skull, and the animal was operated again, sometimes with replacing the cochlear implant electrode. If the skull was too damaged, the animal was sacrificed.

At the end of the 2-hours sisomycin treatment, recordings were made from the implanted cochlear electrode in response to tone pips. No evoked potential could be measured except for some animals from which a small response with a latency of several milliseconds was recorded for the 2-kHz tone pip at 90 dB SPL. However, when adding a continuous contralateral noise at 20 dB below the level of the tone pip to avoid acoustic cross-talk, no response was obtained.

2.3 Stimuli

During the experiment, electrical stimulation was delivered using a stimulus isolator current generator (AM systems model 2200), the output of which was connected through the head connector to the cochlear implant electrode (active) and to one screw in the frontal bone (reference). This generator delivered a current proportional to an arbitrary input voltage waveform. The voltage waveform was acquired at the output of an Advanced Bionics HiRes90k test implant connected to a purely resistive load. This allowed us to use pulse shapes similar to those used in previous human CI studies. The stimuli were designed and controlled using the BEDCS software (Advanced Bionics).

Stimuli were trains of electrical pulses delivered at a rate of 8 Hz. The four pulse shapes illustrated in Figure 1 were tested, including symmetric biphasic (SYM), reversed pseudomonophasic (RPS), reversed pseudomonophasic with an interphase gap of 6.5 ms (RPS-IPG) and triphasic (TRI) pulses. The top pulse shapes are assumed to favor cathodic stimulation, hence the "C" at the end of their acronyms while the bottom pulse shapes are assumed to favor anodic stimulation, hence the "A". The recordings were synchronized to the onset of the cathodic phase for the top pulse shapes and to the onset of the anodic phase for the bottom pulse shapes.

The duration of the phases of SYM, of the short, high-amplitude phases of RPS and RPS-IPG and of the central phase of QP was 86 μs. The long, low-amplitude phases of RPS and RPS-IPG were 8 times longer and lower in amplitude to maintain the total charge balanced. The first and third phases of TRI had a duration of 43 μs and the same amplitude as the central phase.

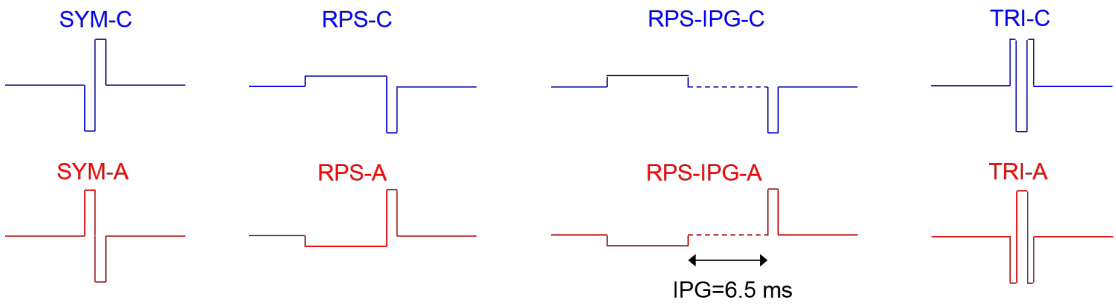


Figure 1. Schematic representation of the electrical pulse shapes used for stimulation, including symmetric biphasic cathodic-first (SYM-C) and anodic-first (SYM-A), Reversed pseudomonophasic with a short, high-amplitude cathodic (RPS-C) or anodic (RPS-A) phase, Reversed pseudomonophasic with a 6.5-ms interphase gap (RPS-IPG-C and RPS-IPG-A) and triphasic with a cathodic (TRI-C) or anodic (TRI-A) central phase.

2.4 Sessions

Over months, several sessions of electrical stimulation from the cochlear implant and recording from the right IC were performed. Each session lasted about one hour during which the animals were sedated. In each session and for each pulse shape, a threshold level was first defined as the level which produced a detectable evoked potential from the IC after averaging; the maximum level of stimulation was defined as that which triggered a detectable motor response of the animal's left cheek. Between these two values, a series of current levels were defined in 1.5-dB steps. For each pulse shape and each level tested, the recordings alternated between polarities. Between 50 and 100 averages were obtained for each combination of pulse shape, polarity and level and the corresponding IC evoked potentials were recorded and kept in computer memory.

3 Results

3.1 Morphology and response amplitude of the IC evoked potential

The bottom traces in Figures 2A and 2B show examples of IC evoked potentials obtained for subject G2 two weeks after deafening in response to RPS-IPG-C and RPS-IPG-A, respectively. For comparison, the top traces in each panel show recordings obtained from the same subject in response to 16-kHz tone pips *before deafening*. The morphology of the response was similar for acoustic and electric stimulation and consisted of a first positive peak (P1) followed by a negative peak (N1) and a second positive peak (P2). As the first peak was not always clearly detectable, the amplitude of the response was arbitrarily defined as the difference in voltage between the negative peak and the second positive peak. The response amplitude varied from about 20 to 600 µV depending on subjects and pulse shapes. The dynamic range, as defined by the difference between the maximum stimulation level and the minimum level that produced a visually detectable response ranged from about 3 to 18 dB, which compare well with usual values of dynamic range for individual CI electrodes.

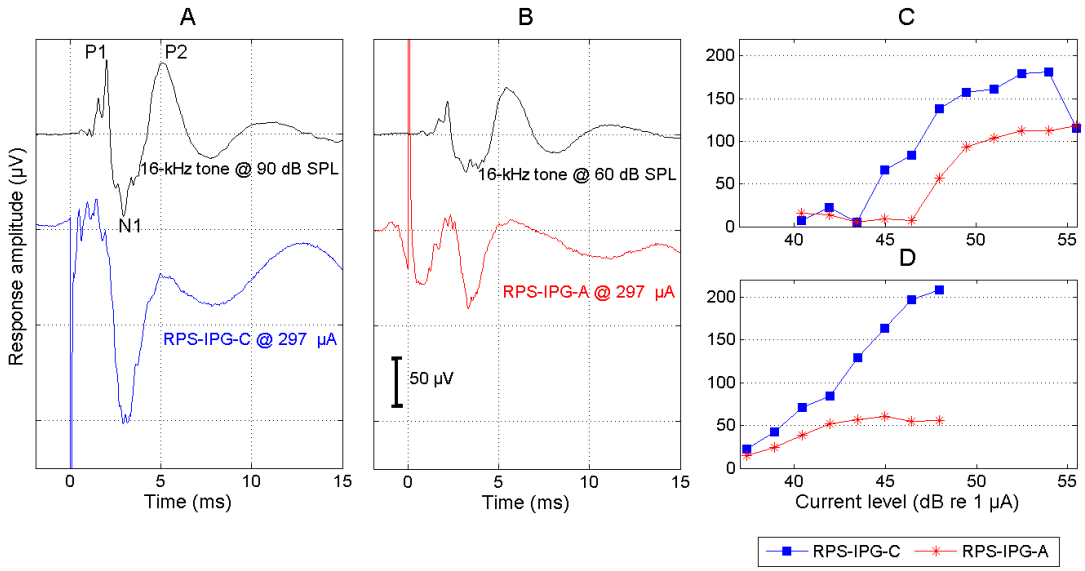


Figure 2. A, Inferior colliculus evoked potential obtained two weeks after deafening for G2 in response to RPS-IPG-C pulses. The top trace shows the response obtained before deafening for an acoustic tone pip presented at 90 dB SPL. B, Same as A for RPS-IPG-A. The top trace shows the response to a 60-dB SPL tone pip. C, Response amplitude (N1-P2) as a function of current level for G2 two weeks after deafening. D, Same as C seven months after deafening.

Figure 2C shows the response amplitude as a function of current level for the same subject and the same pulse shape. It can be seen that at all current levels, the response was larger for the cathodic (RPS-IPG-C) than for the anodic (RPS-IPG-A) pulse. As depicted in the next section, this trend was observed in most of our subjects. Because the RPS-IPG stimulus had a large interphase gap between the two phases, the responses illustrated here presumably represent the responses to each individual polarity. Input/output growth functions from the same animal collected 7 months after deafening are shown in Figure 2D. Although the dynamic range shifted over time, the effect of polarity was very similar to that obtained just after deafening. Matching the response amplitudes obtained for electrical stimulation to those measured acoustically before deafening, it can be deduced that the difference in response amplitude between the two polarities is equivalent to an acoustic level difference of about 30 dB.

Figure 3 shows the amplitude of the response for cathodic stimulation as a function of that for anodic stimulation for the four pulse shapes SYM, RPS, RPS-IPG and TRI and all subjects and levels tested. A very large majority of the data points lie above the diagonal. This illustrates that, for almost all our recordings, so-called cathodic pulses produced a larger response than anodic pulses and that this was the case for all pulse shapes tested. For subjects G8-G10 who had several recording sites, only the recordings from the electrode giving the largest responses are shown and were further analysed. The relative effects between pulse shape and polarity were, however, consistent across the different recording sites.

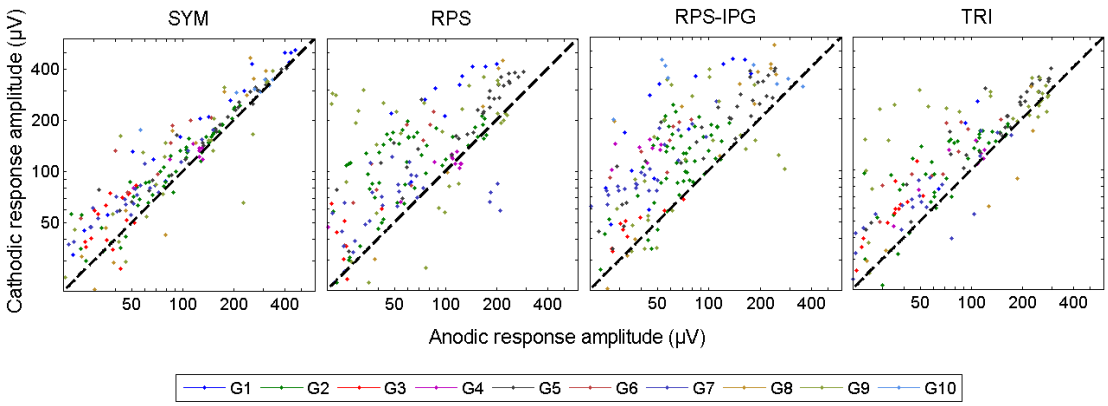


Figure 3. Summary of cathodic and anodic response amplitudes obtained for all sessions and at all levels tested. Each panel shows the results for one of the pulse shapes. The diagonal represents equal sensitivity to anodic and cathodic stimulation.

3.2 Effect of polarity as a function of time

For a given pulse shape and session, the input/output growth functions obtained for the two polarities were analyzed to determine the current levels needed for each of them to reach an arbitrarily-defined IC response amplitude. This arbitrary amplitude was equal to 50% of the maximum response amplitude obtained for this particular pulse shape, subject and session and could, therefore, differ from one session to the next. In some cases, the maximum response amplitude obtained with one polarity was smaller than this arbitrarily-defined amplitude. If this was so, the highest possible response amplitude was taken to perform the comparison without the need to extrapolate the data.

Figure 4A shows the difference in current level between RPS-C and RPS-A when both produced the same response amplitude. This quantity referred to as the "polarity effect" is expressed in dB and is illustrated as a function of time after deafening for the different subjects. Although only the data for RPS are shown, similar results were obtained for RPS-IPG and TRI. If neural degeneration had an effect on polarity sensitivity, we would expect the polarity effect to increase over time and to reach positive values when fibers are degenerated. This does not seem to be the case as the level difference between polarities rather tends to decrease right after deafening for several subjects and then to remain roughly constant.

Given the absence of an effect of time after deafening, the polarity effects obtained in all sessions were averaged for each animal and each pulse shape. As shown by the star symbols in Figure 4B, the magnitude of the polarity effect differed across pulse shapes. The mean across subjects was -1.4, -3.1, -5.3 and -2.3 dB for SYM, RPS, RPS-IPG and TRI, respectively. Paired-sample t-tests with Bonferroni correction showed that this effect was larger for RPS than for SYM ($t(8)=3.6$, $p<0.05$) and also for RPS-IPG than for SYM ($t(9)=3.7$, $p<0.05$). This is consistent with the idea that asymmetric pulses make one polarity dominate over the other more effectively than symmetric pulses. For comparison, filled symbols show the polarity effects obtained using loudness-balancing experiments in human CI listeners for similar pulse shapes (Macherey et al., 2006, 2008; Undurraga et al., 2013; Carlyon et al., 2013). Although the polarity effect is opposite in sign for the two species, it is interesting to note that its magnitude is similar for RPS and for TRI. For RPS-IPG, the smaller effect obtained psychophysically may be due to the fact that the long, low-amplitude phase contributed to perception whereas it did not contribute to the neural responses recorded here.

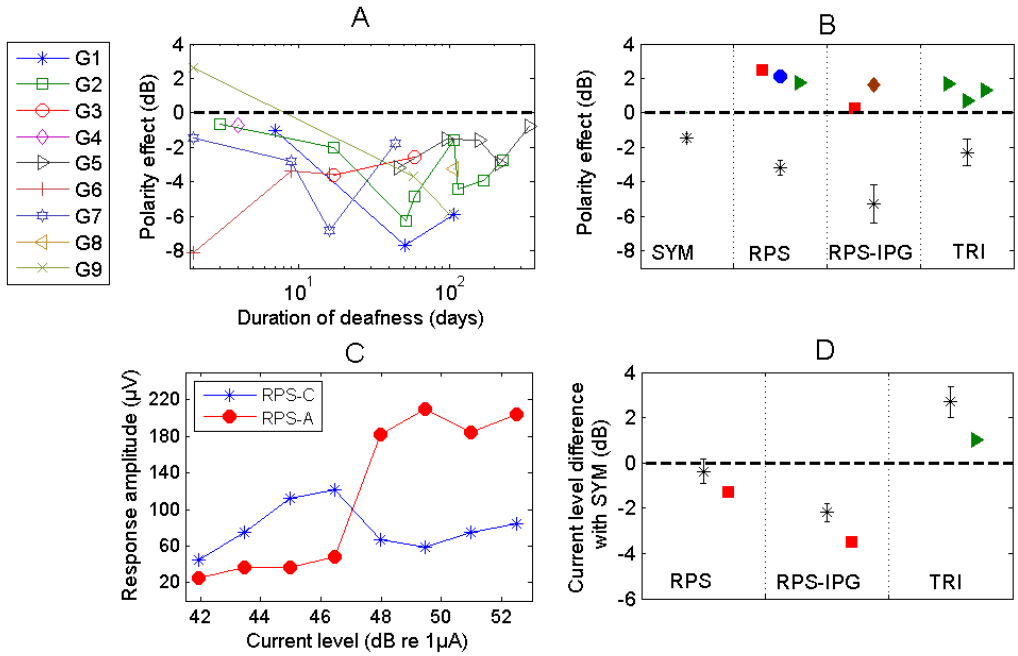


Figure 4. A, Difference in current level between RPS-C and RPS-A required to obtain the same response amplitude in the IC. B, Star symbols: mean polarity effect across guinea-pigs subjects (+/- 1 s.e.). Filled symbols: mean polarity effects obtained at most comfortable loudness for similar pulse shapes in previous human studies. Squares: Macherey et al., 2006; circles: Undurraga et al., 2013; triangles: Carlyon et al., 2013 ; diamonds : Macherey et al., 2008. Human data were obtained with non-reversed versions of RPS and RPS-IPG, meaning that the short, high-amplitude phase was presented first. The three data points for comparison with TRI correspond to the 'TP' condition of Carlyon et al., 2013 for Advanced Bionics (1st point) and Med-EL (2nd point) subjects and to their 'QP' condition for Cochlear subjects (3rd point). C, Example of non-monotonic growth function obtained for RPS in G7. D, Difference in current level between each asymmetric pulse shape and SYM needed to obtain the same response amplitude. Star symbols: results of the present study averaged across subjects (+/- 1 s.e.). Filled symbols show corresponding human psychophysical data measured at most comfortable loudness.

Closer inspection of the data revealed that the input/output growth functions sometimes had non-monotonic shapes. An example obtained for G7 one month after deafening is illustrated in Figure 4C for RPS-A and RPS-C. The function for RPS-C shows, at some point in the dynamic range, a decrease in response amplitude with increases in level. In this case, the cathodic pulse was more effective than the anodic pulse at low levels but less effective at high levels. Such non-monotonicities were observed in seven of the ten subjects but only in about 13% of the cathodic and 4% of the anodic functions. Furthermore, they did not always occur for the same pulse shape across sessions.

3.3 Effect of pulse shape

The difference in level between SYM-C and the various asymmetric pulse shapes RPS-C, RPS-IPG-C and TRI-C when they all produced the same response amplitude was determined. The response amplitude at which the pulse shapes were compared was equal to 50% of the maximum amplitude of SYM-C. Given there was no effect of duration of deafness on the magnitude of these effects, the data were averaged across sessions for each animal and analyzed in a one-way repeated measures ANOVA. The star symbols in Figure 4D show the mean *across* animal subjects. The effect of pulse shape was found to be highly significant ($F(3,18)=22.76$, $p<0.001$). Pairwise comparisons showed that, in order to reach the same response amplitude, RPS-IPG required 2.2 dB less current than SYM ($p<0.05$) while TRI required 2.7 dB more. There was no significant difference between SYM and RPS. For comparison, the filled symbols show the

effect of pulse shape as observed in human CI listeners. For these data, however, the level difference was obtained using loudness balancing between *anodic* pulse shapes. Here again, the magnitude of the effect of pulse shape is similar in both species.

4 Discussion

All guinea-pigs tested in the present study showed higher sensitivity to cathodic stimulation. This effect of polarity was consistent across all levels tested, ranging from threshold to the maximum level. This corroborates the results obtained in most previous animal experiments performed in cats and guinea pigs except Miller et al.'s study (1998) remains to date the only animal study showing greater sensitivity to anodic stimulation in auditory nerve stimulation.

Duration of deafness did not influence this result. This suggests that neural degeneration may not *entirely* account for the higher human sensitivity to anodic stimulation. However, the morphology of the nerve fibers and the cochlear geometry are different in both species. In particular, the cell body of the spiral ganglion neurons is larger and unmyelinated in humans and this difference may play an additional role on neural excitation (Rattay et al., 2001). One limitation of the present study was that we did not perform histology on the subjects' cochleae and that we did not know the exact state of degeneration at each session.

In a few cases, input/output growth functions were non-monotonic, showing a decrease in response amplitude with increases in level. This occurred in a few instances in 7 of the 10 subjects tested but was not always specific to the cathodic stimuli. Similar non-monotonic functions have already been observed at the level of the auditory nerve (Hartmann et al., 1984; Miller et al., 1998). They may reflect the blocking of action potentials along the nerve fibers (Frijns et al., 1996) and, therefore, remain a possible candidate explanation for the higher efficiency of anodic stimulation at suprathreshold levels in humans. The fact that this non-monotonic behavior was observed only in some sessions may reveal a dependence on the electrode position which might have changed over time. This remains, however, speculative as we did not control for electrode positioning across sessions.

Finally, the difference in efficiency between the various pulse shapes was consistent with results obtained in previous animal and human experiments (Miller et al., 2001; Macherey et al., 2006), showing that pseudomonophasic pulses with or without a gap between the two phases require less current than symmetric pulses to elicit the same response but that triphasic pulses require *higher* current levels.

Acknowledgments

The authors acknowledge funding from the Agence Nationale de la Recherche (grant numbers ANR-11-PDOC-0022 and ANR-TECSAN-040-STESU).

References

Arnold, S., & Burkard, R. (1998). The auditory evoked potential difference tone and cubic difference tone measured from the inferior colliculus of the chinchilla. *J Acoust Soc Am*, 104(3 Pt 1):1565-1573.

Carlyon, R.P., Deeks, J.M., & Macherey, O. (2013). Polarity effects on place pitch and loudness for three cochlear-implant designs and at different cochlear sites. *J Acoust Soc Am*, 134(1):503-509.

Cazals, Y., Aran, J.M., & Charlet de Sauvage, R. (1983). Artificial activation and degeneration of the cochlear nerve in guinea pigs. *Arch Otorhinolaryngol*, 238(1):1-8.

Deng, A., Lu, J., & Sun, W. (2010). Temporal processing in inferior colliculus and auditory cortex affected by high doses of salicylate. *Brain Res*, 1344:96-103.

Fayad, J.N., & Linthicum, F.H.Jr. (2006). Multichannel cochlear implants: relation to histopathology to performance. *Laryngoscope*, 116(8):1310-1320.

Frijns, J.H., de Snoo, S.L., & ten Kate, J.H. (1996). Spatial selectivity in a rotationally symmetric model of the electrically stimulated cochlea. *Hear Res*, 95(1-2):33-48.

Hartmann, R., Topp, G., & Kinke, R. (1984). Discharge patterns of cat primary auditory fibers with electrical stimulation of the cochlea. *Hear Res*, 13(1):47-62.

Klop, W.M., Hartlooper, A., Briaire, J.J., & Frijns, J.H. (2004). A new method for dealing with the stimulus artefact in electrically evoked compound action potential measurements. *Acta Otolaryngol*, 124(2):137-143.

Koitchev, K., Guilhaume, A., Cazals, Y., & Aran, J.M. (1982). Spiral ganglion changes after massive aminoglycoside treatment in the guinea pig. Counts and ultrastructure. *Acta Otolaryngol*, 94(5-6):431-8.

Leake, P.A., & Hradek, G.T. (1988). Cochlear pathology of longterm neomycin induced deafness in cats. *Hear Res*, 33(1):11-33.

Macherey, O., van Wieringen, A., Carlyon, R.P., Deeks, J.M., & Wouters, J. (2006). Asymmetric pulses in cochlear implants: effects of pulse shape, polarity, and rate. *J Assoc Res Otolaryngol*, 7(3):254-266.

Macherey, O., Carlyon, R.P., van Wieringen, A., Deeks, J.M., & Wouters, J. (2008). Higher sensitivity of human auditory nerve fibers to positive electrical currents. *J Assoc Res Otolaryngol*, 9(2):241-251.

Macherey O., Deeks, J.M., & Carlyon, R.P. (2011). Extending the limits of place and temporal pitch perception in cochlear implant users. *J Assoc Res Otolaryngol*, 12(2):233-251.

Miller, C.A., Abbas, P.J., Rubinstein, J.T., Robinson, B.K., Matsuoka, A.J., & Woodworth, G. (1998). "Electrically evoked compound action potentials of guinea pig and cat: responses to monopolar, monophasic stimulation. *Hear Res*, 119(1-2):142-154.

Miller, C.A., Abbas, P.J., Robinson, B.K., Rubinstein, J.T., & Matsuoka, A.J. (1999). Electrically evoked single-fiber action potentials from cat: responses to monopolar, monophasic stimulation. *Hear Res*, 130(1-2):197-218.

Popelár, J., Erre, J.P., Aran, J.M., Cazals, Y. (1994). Plastic changes in ipsi-contralateral differences of auditory cortex and inferior colliculus evoked potentials after injury to one ear in the adult guinea pig. *Hear Res*, 72(1-2):125-34.

Rattay, F. (1989). Analysis of models for extracellular fiber stimulation. *IEEE Trans Biomed Eng*, 36(7):676-682.

Rattay, F., Lutter, P., ad Felix, H. (2001). A model of the electrically excited human cochlear neuron. I. Contribution of neural substructures to the generation and propagation of spikes. *Hear Res*, 153(1-2):43-63.

Undurraga, J.A., van Wieringen, A., Carlyon, R.P., Macherey, O., & Wouters, J. (2010). Polarity effects on neural responses of the electrically stimulated auditory nerve at different cochlear sites. *Hear Res*, 269(1-2):146-161.

Undurraga, J.A., Carlyon, R.P., Wouters, J., & van Wieringen, A. (2013). The polarity sensitivity of the electrically stimulated human auditory nerve measured at the level of the brainstem. *J Assoc Res Otolaryngol*, 14(3):359-377.

Chapter 16

Assessing the Firing Properties of the Electrically Stimulated Auditory Nerve using a Convolution Model

Stefan B. Strahl[1], Dyan Ramekers[2,3], Marjolijn M.B. Nagelkerke[2], Konrad E. Schwarz[1], Philipp Spitzer[1], Sjaak F.L. Klis[2,3], Wilko Grolman[2,3], Huib Versnel[2,3]

[1] *R&D MED-EL GmbH, Innsbruck, Austria*
[2] *Department of Otorhinolaryngology and Head & Neck Surgery, University Medical Center Utrecht, Utrecht, Netherlands*
[3] *Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, Netherlands*

Contact author: SBS, stefan.strahl@medel.com

Abstract

The electrically evoked compound action potential (eCAP) is a routinely performed electrophysiological measure of the auditory nerve in cochlear implant users. Using a convolution model of the eCAP, additional information about the firing properties of the auditory nerve can be obtained. In this study, eCAPs for a monopolar biphasic pulse were analysed in two groups: acutely implanted normal-hearing and deafened guinea pigs, and human cochlear implant recipients. The estimated firing probability of the auditory nerve could be parameterised for most eCAPs by a bimodal probability distribution consisting of two Gaussian distributions with an average latency difference of 0.4-0.5 ms. The ratio of the late and early Gaussian distributions increased with neural degeneration in the guinea pig and this ratio decreased with stimulation intensity. This decrease with intensity was also observed in humans. The latency of the early Gaussian distribution decreased with neural degeneration in the guinea pig. Indirectly, this was observed in humans as well, assuming that the cochlear base exhibits more neural degeneration than the apex. Differences between guinea pigs and humans were observed, among other parameters, in the width of the early component: very robust in guinea pig, and dependent on stimulation intensity and cochlear region in humans. We conclude that the deconvolution of the eCAP is a valuable addition to existing eCAP analyses, in particular in revealing two separate firing components.

Keywords: auditory nerve, eCAP, cochlear implant, deconvolution, firing probability, neural health

1 Introduction

Most cochlear implant (CI) systems allow for the recording of the auditory nerve's response to an electric stimulus - the electrically evoked compound action potential (eCAP). This objective measure is an important clinical tool to assess the electrode-nerve interface of a CI recipient (Miller et al., 2008). Routinely, the lowest stimulation level that evokes an eCAP ('threshold') is determined. Additionally, the eCAP amplitude as a function of stimulation intensity, the input/output function, is often recorded and quantified by its slope and maximal observed eCAP amplitude. The morphology of the eCAP waveform is usually not evaluated.

1.1 Mathematical Model of the Compound Action Potential

Most models of the evoked response of the auditory nerve are based on the mathematical concept of a convolution (Goldstein & Kiang, 1958, see also Figure 1). The recorded compound action potential (*CAP*) is described as the convolution of the unit response (*UR*, i.e., the waveform resulting from a single action potential) with the compound discharge latency distribution (*CDLD*), the sum of the responses of all individual nerve fibres.

$$\mathrm{CAP}(t) = \int_{-\infty}^{t} \mathrm{CDLD}(\tau)\mathrm{UR}(t-\tau)d\tau \qquad (1)$$

This mathematical model was validated with simultaneous recordings of acoustically evoked CAPs from the round window and single-fibre responses from the auditory nerve in guinea pigs (Versnel et al., 1992b). The single-fibre responses were used to estimate the CDLD. The unit response was recorded as well, showing only slight changes in noise-damaged cochleas (Versnel et al., 1992a).

Most mathematical models of the CAP are concerned with solving the forward problem, i.e. predicting the CAP by modelling the activation of single nerve fibres and assuming a convolution with the unit response (Teas et al., 1962; de Boer, 1975; Elberling, 1976b; Kiang et al., 1976; Dolan et al., 1983; Versnel et al. 1992 b; Frijns et al., 1996; Miller et al., 1999; McMahon & Patuzzi, 2002; Briaire & Frijns, 2005; Whiten, 2007).

A few studies addressed the inverse problem, i.e. predicting the firing probability by deconvolution of acoustically evoked CAPs with a unit response (Elberling, 1976a; Charlet de Sauvage et al., 1987). To our knowledge no study investigated the inverse problem for eCAPs. In this study, we perform a deconvolution of eCAPs to examine CDLDs, presuming that they could provide information about the neural status of the auditory periphery. It has been shown that the eCAP in a guinea pig exhibits differences in its morphology depending on the neural health of the auditory nerve (Ramekers et al., 2014, 2015).

2 Methods

2.1 eCAP Recording System

eCAP measurements were performed with MED-EL PULSAR or SONATA cochlear implants (for details see Neustetter et al., 2012). A charge-balanced biphasic pulse with a variable inter-phase gap (IPG) and a phase duration of 30 µs was presented at an intra-cochlear stimulation electrode in a monopolar configuration. The stimulation amplitude was defined in current unit (cu), where 1 cu corresponds to approximately 1 µA. The eCAP was recorded on a neighbouring electrode with a minimum delay of 125 µs from stimulus onset to reduce stimulation artefacts. The recording window was 1.7 ms. Single eCAP recordings were separated by at least 20 ms, assuming a relative refractory time below 10 ms (Brill et al., 2009; Ramekers et al., 2015). Any stimulation or recording artefact was removed using an alternating polarity and zero amplitude template paradigm, respectively (for details see Brill et al., 2009).

2.2 eCAP Data Sets

eCAPs obtained from a guinea pig model of deafness were analysed to assess any effects of the neural status of the auditory nerve on its firing probability. The second data set consisted of eCAPs from human CI users to evaluate if the in the animal model observed effects are comparable in humans.

Guinea Pig Deafness Model eCAP Dataset

Data were acquired from 18 guinea pigs during an extensive series of experiments (Ramekers et al., 2014, 2015). These animals were divided into three groups of 6 animals: normal-hearing

(NH), 2 weeks deaf (2WD) and 6 weeks deaf (6WD). Methods of deafening and eCAP recordings are described in detail in Ramekers et al. (2014).

Briefly, profound cochlear hair cell loss was induced by co-treatment of kanamycin and furosemide. It resulted in mild degeneration of the auditory nerve after 2 weeks (25% loss of spiral ganglion cells, SGCs) and severe degeneration after 6 weeks (60% loss of SGCs).

The eCAP recordings were performed in isoflurane-anaesthetized animals. eCAPs were evoked and recorded with two electrodes placed in the basal turn through a cochleostomy. Stimuli consisted of biphasic pulses with a phase duration of 30 μs; two IPGs were applied: 2.1 or 30 μs. In this study, three different stimulation intensities were analysed: just above threshold (first visible eCAP), halfway the input/output function ('intermediate'), and at maximum stimulation intensity typically corresponding to saturation (800 cu). Waveforms were averaged over 900 iterations.

Human Clinical eCAP Dataset

From a multicentre study (Senn et al., 2012) eCAPs recorded post-operatively with the clinical system software MAESTRO (MED-EL GmbH, Innsbruck) in 52 awake human subjects (25 female, 27 male) were selected. The selection criteria were the availability of a pre-operative audiogram and eCAPs at maximal stimulation level of at least 100 μV amplitude. The mean age at onset of hearing loss was 19 years with a range of 0 to 72 years. The mean age at implantation was 46 years with a range of 15 to 79 years.

The recordings were manually analysed by three experts. Similar to the animal model, three different stimulation levels were selected: the lowest stimulation level ('threshold') that evoked an eCAP, the stimulation level ('maximum') resulting in the largest eCAP response typically corresponding to the loudest acceptable presentation level and a stimulation level ('intermediate') corresponding to about 50% eCAP amplitude. The eCAPs were obtained from stimulation electrodes in the apical (electrode contact 2), middle (electrode contacts 4 or 5) and basal region (electrode contacts 9 or 10) of the cochlea. Waveforms were averaged over 25 iterations and 5 kHz low-pass filtered (fifth-order Butterworth) to remove any remaining recording noise. The IPG was kept constant at 2.1 μs.

2.3 Deconvolution of the eCAP

This study examines the inverse problem where the eCAP waveform is known and the CDLD of the auditory nerve has to be determined. Following Eq. (1) and (Versnel et al. 1992a) the unit response UR can be assumed constant and was modelled by Equations 2a and 2b with U_N = 0.12 μV, σ_N = 0.12 ms describing the negative and U_P = 0.045 μV, σ_P = 0.16 ms describing the positive part. The cross point is defined with t_0 = -0.06 ms (Figure 1A).

$$UR(t) = \frac{U_N}{\sigma_N}(t - t_0)e^{\frac{1}{2} - \frac{(t-t_0)^2}{2\sigma_N{}^2}}, t < t_0 \quad \text{(2a)}$$

$$UR(t) = \frac{U_P}{\sigma_P}(t - t_0)e^{\frac{1}{2} - \frac{(t-t_0)^2}{2\sigma_P{}^2}}, t \geq t_0 \quad \text{(2b)}$$

Assuming the same unit response for all contributing fibres the CDLD can be obtained directly from the recorded eCAP by deconvolution. The deconvolution was performed in the frequency domain with

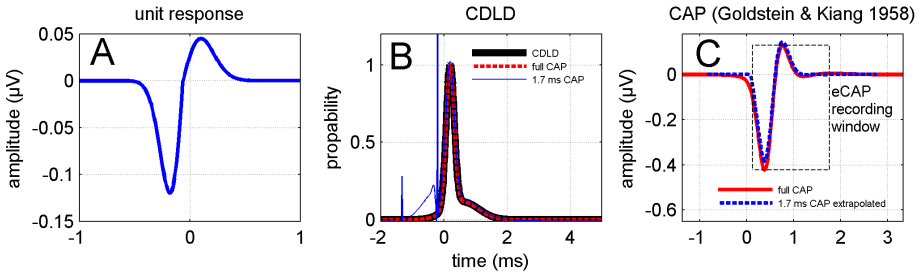


Figure 1. The assumed elementary unit response of a single nerve fibre is shown in **A**. Its convolution with a CDLD (**B**, thick black line) results in the CAP waveform shown in **C** (red line). The monotonic extrapolation of the time-limited eCAP recording is shown in **C** (blue dashed line). The derived CDLDs obtained by Equation 3 are shown in **B** for the full eCAP (red dashed line) and a simulation of the time-limited and extrapolated eCAP (thin blue line).

$$CDLD(t) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(eCAP(t))}{\mathcal{F}(UR(t))}\right) \qquad (3)$$

where $\mathcal{F}$ represents the Fourier transform and $\mathcal{F}^{-1}$ the inverse Fourier transform.

Current CI systems cannot record the eCAP directly from stimulus onset on due to an amplifier saturation caused by the stimulation pulse. Resulting stimulus artefacts around onset were removed. The baseline was estimated over the final 200 µs and, assuming that the eCAP decays monotonically to baseline before and after the recording window, the eCAPs were linearly extrapolated to baseline (Figure 1C) before performing the deconvolution. The extrapolation only affected the CDLD before stimulus onset (see Figure 1B). The deconvolution of the eCAP can be understood as applying an inverse low-pass filter, amplifying any unwanted high-frequency noise. Therefore, the extrapolated eCAPs were low-pass filtered by a 50-point moving-average filter applied twice (once in forward and once in reverse direction) to achieve a zero-phase digital filtering. Any remaining high-frequency noise in the CDLD was removed by a 2.5 kHz fifth-order Butterworth low-pass filter.

2.4 Parametrisation of the CDLD

The CDLDs derived from eCAP responses from both data sets exhibited a skewed, quasi-bimodal, distribution (see Figure 2). To perform further analysis, a parametrisation of the CDLD was performed by a two-component Gaussian mixture model (GMM).

$$CDLD = a_1\,\mathcal{N}(\mu_1, \sigma_1) + a_2\,\mathcal{N}(\mu_2, \sigma_2) \qquad (4)$$

The GMM was fitted to the CDLDs using a nonlinear least-squares regression (nlinfit, MATLAB; Mathworks, Natick, MA, USA) with initial start values being manually optimised if needed to achieve an adequate fit ($R^2 > 0.93$). As outcome parameters we considered μ_1 and μ_2 (corresponding to the peak latencies), σ_1 and σ_2 (reflecting peak width), and the ratio of the components a_2/a_1.

3 Results

3.1 Guinea Pig Deafness Model eCAPs

Figure 2 shows eCAP examples for an NH (A), 2WD (B) and 6WD animal (C). In each animal the deconvolution of the recorded eCAPs resulted in CDLDs with a fast rising peak, typically narrow (<0.5 ms width), in most cases followed by a shoulder-shaped component or a second peak (Fig. 2A-C, bottom row). In those cases the CDLD could be well fitted with the two-component GMM (Eq. 4), remaining single-peak cases (26% of all CDLDs) could be well fitted with a single

Gaussian. Figure 3 and Table 1 show group averages of CDLD parameters. Two variables significantly varied with group (i.e., with degeneration of the auditory nerve): μ_1 and a_2/a_1 ($p<0.01$; rmANOVA with IPG and stimulation intensity as within factors, and group as between factor). With more nerve degeneration (6WD vs. 2WD and NH) the peak latency μ_1 was shorter and a_2/a_1 was larger.

A larger IPG is more effective resulting in larger eCAPs, and correspondingly, most eCAP and CDLD variables significantly depended on IPG. The effects with group described above were similar for both IPGs used (2.1 and 30 μs). However, a significant decrease a_2/a_1 with increasing stimulation intensity is only found for an IPG of 30 μs ($p<0.01$; rmANOVA with stimulation intensity as within factor and group as between factor).

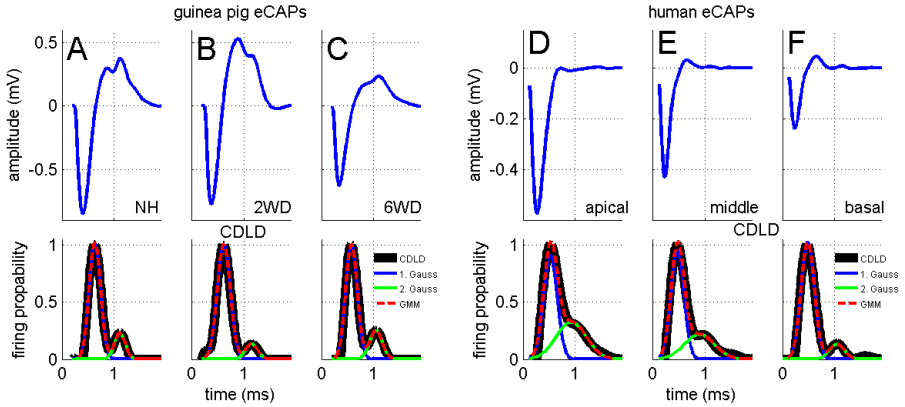


Figure 2. Example eCAPs for maximum stimulation level (upper row) and the corresponding CDLDs and GMM fits (bottom row). The left three panels are from the three guinea pig groups recorded at 30 μs IPG. The right three panels are from one human subject from each of the three cochlear regions in one human subject recorded at 2.1 μs IPG.

3.2 Human Clinical eCAPs

Figure 2 shows examples of eCAPs and corresponding CDLDs for an apical (D), middle (E) and basal (F) stimulation site in one human subject. Human CDLDs showed a morphology comparable to the guinea pig data, and in a small subgroup (15%) also uni-modal firing probability functions were observed. Subjects were divided into similar-sized groups based on their pre-operative pure tone average (PTA): ≤105 dB HL (24 subjects) or >105 dB HL (29 subjects) – assuming that any residual hearing could correlate with the neural health of the auditory nerve. Figure 3 and Table 1 show averages of CDLD parameters for different cochlear regions. No significant between-subject effects were found ($p>0.05$; rmANOVA with cochlear region and stimulation intensity as within factors, and group as between factor). A significant decrease of a_2/a_1 with increasing stimulation intensity ($p<0.001$) and distance from the apex (apex vs. middle and base, $p<0.05$) was found. The latency μ_1 was significantly longer in the apical cochlear region compared to middle and base ($p<0.001$). The width of the first Gaussian σ_1 was significantly smaller in the apex vs. middle and base ($p<0.05$) and with increasing stimulation intensity ($p<0.001$).

Table 1. CDLD parameters for guinea pig groups (upper section) and humans grouped by cochlear region (lower section) averaged over stimulation level and IPG.

		μ_1 (ms)	μ_2-μ_1(ms)	σ_1 (ms)	σ_2 (ms)	a_2/a_1
Guinea Pig	*NH*	0.59	0.56	0.11	0.11	0.15
	2WD	0.61	0.50	0.12	0.10	0.14
	6WD	0.55	0.49	0.11	0.14	0.28
Human	*Apex*	0.54	0.42	0.13	0.27	0.96
	Middle	0.50	0.42	0.14	0.27	0.86
	Base	0.51	0.46	0.14	0.26	0.85

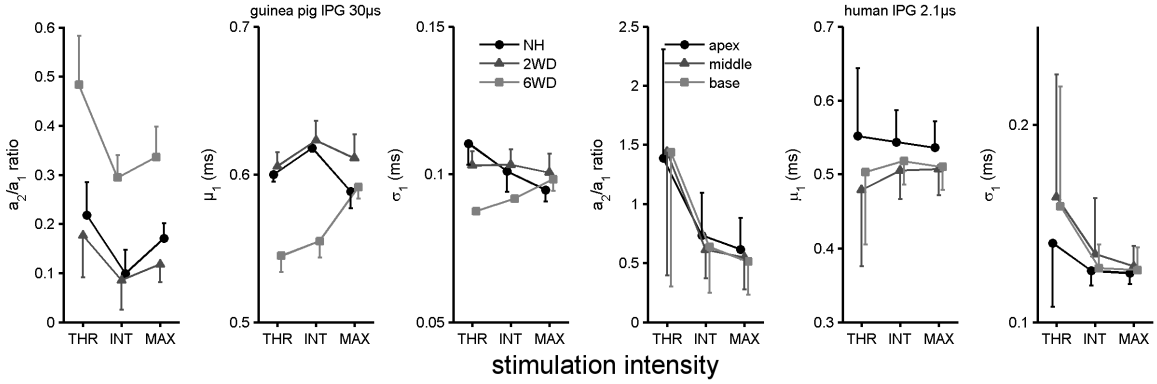


Figure 3. CDLD parameters a_2/a_1, μ_1 and σ_1 shown for guinea pigs (left) and humans (right). THR, threshold; INT, intermediate; MAX, maximum. Errorbars are SEM.

4 Discussion

4.1 Comparing Deconvolution Results from Guinea Pigs and Humans

We observed the following similarities between the CDLDs resulting from the eCAP deconvolutions in guinea pigs and humans. First, the CDLDs could be described by the sum of two Gaussian distributions (Fig. 2). Second, the peak latency of the early component decreased with increasing neural degeneration, assuming for the human data more degeneration at the cochlear base than near the apex (Fig. 3, μ_1). Third, the ratio of the late and early component decreases with stimulation intensity (Fig. 3, a_2/a_1). Notable differences were observed in particular with respect to the width of the early component, which was very robust in guinea pigs, and quite variable with stimulation intensity and cochlear region in humans.

4.2 Choice of UR Waveform

The main challenge of the eCAP deconvolution is the dependence of the CDLDs on the choice of UR. For instance, with a larger U_P a_2/a_1 increases since the P_2 component of eCAP can be reduced by increasing the N_2. We have chosen our parameters based on experimental estimates of the UR in guinea pigs (Prijs, 1986; Versnel et al. 1992 a). Therefore, we can be fairly confident about our choice for the eCAP deconvolutions in the guinea pig. The UR in humans is thought to be similar (Briaire & Frijns, 2005; Whiten, 2007). However, the cochlear morphology is quite different in humans. The UR may change with stimulation site in the human cochlea and explain in part the cochlear region dependent variation of the CDLD seen in the human dataset. Also our assumption of the same UR across the nerve population might be too coarse (Westen et al., 2011), in particular considering effects of degeneration as demyelination. To further address this point, a sensitivity analysis of the convolution model is needed.

4.3 Parametrisation of the CDLD

Single-fibre responses to electric pulses (estimated by poststimulus time histograms) showing a fast rise and a slow decline (van den Honert & Stypulkowski, 1984; Sly et al., 2007) may be modelled best with a gamma function (Versnel et al. 1992 b). However, the distribution of response latencies between fibres may be Gaussian. Therefore, we tested fits of CDLDs with gamma functions and with Gaussian functions, and both in guinea pigs and humans we observed superior fits with Gaussian functions. Due to the large population size and assuming that the discharge jitter is smaller within than between fibres, the resulting convolution of the gamma function with the Gaussian distribution could explain the Gaussian-like CDLD.

4.4 Interpretation of CDLD Parameters

In most cases a_2/a_1 is significant, in both guinea pigs and humans. This suggests either two subpopulations of fibres and/or neural elements (peripheral vs. central site of excitation) with different latencies and proportions or repeated firings of the same neurons. The latter option is supported by the following arguments (see also in Ramekers et al., 2015). First, the interval between the two components, μ_2-μ_1, is about 0.5 ms, which is around the absolute refractory period estimated from masker-probe responses. Second, the N_2-N_1 interval of the eCAP waveform has been found to correlate with recovery measures (Ramekers et al., 2015).

Alternatively, a reduced relative contribution of the second component with increasing stimulation intensity and more basal position within the cochlea could be explained by the former option. The late component can be ascribed to excitation peripheral to the cell body (Stypulkowski & van den Honert, 1984) which is thought to occur to a larger extent near threshold (Briaire & Frijns, 2005).

5 Conclusion

Obviously, there are multiple differences in experimental settings and biology between the guinea pig and human, which may contribute to different CDLD outcomes. Among others, cause of deafness, cochlear anatomy, duration of deafness, and rate of neural degeneration differ, and there is awake vs. anaesthetized state, chronic vs. acute cochlear implantation, and long vs. short electrode array. Taking these differences into account, a few characteristics appeared to be quite robust across species and experimental setting. Importantly, the deconvolution of the eCAP revealed two separate firing components, which could not easily be detected from the eCAP waveforms themselves. The ratio of the components and the latency of the early component came out as potential markers to assess the condition of the auditory nerve.

References

Briaire, J. J., & Frijns, J. H. M. (2005). Unraveling the electrically evoked compound action potential. *Hearing Research*, *205*(1-2), 143–156.

Brill, S., Müller, J., Hagen, R., Möltner, A., Brockmeier, S.-J., Stark, T., … Strahl, S. (2009). Site of cochlear stimulation and its effect on electrically evoked compound action potentials using the MED-EL standard electrode array. *BioMedical Engineering OnLine*, *8*(1), 40.

Charlet de Sauvage, R., Aran, J. M., & Erre, J. P. (1987). Mathematical analysis of VIIIth nerve cap with a linearly-fitted experimental unit response. *Hearing Research*, *29*(2-3), 105–115.

De Boer, E. (1975). Synthetic whole-nerve action potentials for the cat. *The Journal of the Acoustical Society of America*, *58*(5), 1030–1045.

Dolan, D. F., Teas, D. C., & Walton, J. P. (1983). Relation between discharges in auditory nerve fibers and the whole-nerve response shown by forward masking: an empirical model for the AP. *The Journal of the Acoustical Society of America*, *73*(2), 580–591.

Elberling, C. (1976a). Deconvolution of action potentials recorded from the ear canal in man. *Disorders of Auditory Function II. Academic Press, London*.

Elberling, C. (1976b). Simulation of cochlear action potentials recorded from the ear canal in man. In *Electrocochleography* (pp. 151–168). University Park Press Baltimore.

Frijns, J. H. M., Schoonhoven, R., & Grote, J. J. (1996). The influence of stimulus intensity on spike timing and the compound action potential in the electrically stimulated cochlea: a model study. In *Proceedings of the 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine* (Vol. 1, pp. 327–328 vol.1).

Goldstein, J. M. H., & Kiang, N. Y.-S. (1958). Synchrony of Neural Activity in Electric Responses Evoked by Transient Acoustic Stimuli. *The Journal of the Acoustical Society of America*, *30*(2), 107–114.

Kiang, N. Y. S., Moxon, E. C., & Kahn, A. R. (1976). The relationship of gross potentials recorded from the cochlea to single unit activity in the auditory nerve. In *Electrocochleography* (pp. 95–115). University Park Press Baltimore.

McMahon, C. M., & Patuzzi, R. B. (2002). The origin of the 900 Hz spectral peak in spontaneous and sound-evoked round-window electrical activity. *Hearing Research*, *173*(1-2), 134–152.

Miller, C. A., Abbas, P. J., & Rubinstein, J. T. (1999). An empirically based model of the electrically evoked compound action potential1. *Hearing Research*, *135*(1–2), 1–18.

Miller, C. A., Brown, C. J., Abbas, P. J., & Chi, S.-L. (2008). The clinical application of potentials evoked from the peripheral auditory system. *Hearing Research*, *242*(1–2), 184–197.

Neustetter, C., Zangerl, M., Spitzer, P., & Zierhofer, C. (2012). In-vitro characterization of a cochlear implant system for recording of evoked compound action potentials. *BioMedical Engineering OnLine*, *11*(1), 22.

Prijs, V. F. (1986). Single-unit response at the round window of the guinea pig. *Hearing Research*, *21*(2), 127–133.

Ramekers, D., Versnel, H., Strahl, S. B., Klis, S. F. L., & Grolman, W. (2015). Recovery characteristics of the electrically stimulated auditory nerve in deafened guinea pigs: Relation to neuronal status. *Hearing Research*, *321*, 12–24.

Ramekers, D., Versnel, H., Strahl, S. B., Smeets, E. M., Klis, S. F. L., & Grolman, W. (2014). Auditory-Nerve Responses to Varied Inter-Phase Gap and Phase Duration of the Electric Pulse Stimulus as Predictors for Neuronal Degeneration. *Journal of the Association for Research in Otolaryngology*, *15*(2), 187–202.

Senn, P., van de Heyning, P., L. Arauz, S., Atlas, M., Baumgartner, W.-D., Caversaccio, M., … Skarzynski, H. (2012). Electrically Evoked Compound Action Potentials In Patients Supplied With CI. Presented at the 12th International Conference on Cochlear Implants and Other Implantable Auditory Technologies, Baltimore.

Sly, D. J., Heffer, L. F., White, M. W., Shepherd, R. K., Birch, M. G. J., Minter, R. L., … O'Leary, S. J. (2007). Deafness alters auditory nerve fibre responses to cochlear implant stimulation. *The European Journal of Neuroscience*, *26*(2), 510–522.

Stypulkowski, P. H., & van den Honert, C. (1984). Physiological properties of the electrically stimulated auditory nerve. I. Compound action potential recordings. *Hearing Research*, *14*(3), 205–223.

Teas, D. C., Eldridge, D. H., & Davis, H. (1962). Cochlear responses to acoustic transients: An interpretation of whole-nerve action potentials. *Journal of the Acoustical Society of America*, *34*(9, Pt. II), 1438–1459.

Van den Honert, C., & Stypulkowski, P. H. (1984). Physiological properties of the electrically stimulated auditory nerve. II. Single fiber recordings. *Hearing Research*, *14*(3), 225–243.

Versnel, H., Prijs, V. F., & Schoonhoven, R. (1992a). Round-window recorded potential of single-fibre discharge (unit response) in normal and noise-damaged cochleas. *Hearing Research*, *59*(2), 157–170.

Versnel, H., Schoonhoven, R., & Prijs, V. F. (1992b). Single-fibre and whole-nerve responses to clicks as a function of sound intensity in the guinea pig. *Hearing Research*, *59*(2), 138–156.

Westen, A. A., Dekker, D. M. T., Briaire, J. J., & Frijns, J. H. M. (2011). Stimulus level effects on neural excitation and eCAP amplitude. *Hearing Research*, *280*(1-2), 166–176.

Whiten, D. M. (2007). *Electro-anatomical models of the cochlear implant* (Thesis). Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/38518. Accessed 1 April 2015

Chapter 17

Modeling the Individual Variability of Loudness Perception with a Multi-Category Psychometric Function

Andrea C. Trevino, Walt Jesteadt, Stephen T. Neely
Boys Town National Research Hospital, Omaha, NE, USA

Contact author: A.C.T., Andrea.Trevino@boystown.org

Abstract

Loudness is a measure of suprathreshold perception that provides insight into the status of the entire auditory pathway; cochlear damage manifests primarily as loudness recruitment, while central adaptations may affect the level at which listeners report discomfort. In addition to these broad effects, individuals with matched thresholds show significant individual variability in the details of their loudness perception. As a means to analyze and model listener variability, we introduce the *multi-category psychometric function* (MCPF), a novel representation for categorical data that fully describes the probabilistic relationship between stimulus level and categorical-loudness perception. We present results based on categorical loudness scaling (CLS) data for 15 normal-hearing (NH) adults and 22 adults with hearing loss (HL). A principle-component analysis of the parameterized listener data shows the major underlying factors that contribute to individual variability across listeners. We show how the MCPF can be used to improve CLS estimates, by combining listener models with maximum-likelihood estimation. The MCPF could be additionally used in an entropy-based stimulus-selection technique. These techniques utilize the probabilistic nature of categorical perception, a new dimension of loudness information, to improve the quality of loudness measurements.

1 Introduction

Loudness is a manifestation of nonlinear suprathreshold perception that is altered when cochlear damage exists (Allen, 2008). There are a number of techniques for measuring loudness perception (Florentine, 2011); in this work we will focus on Categorical Loudness Scaling (CLS). CLS is a task that has a well-studied relationship with hearing loss, uses category labels that are ecologically valid, can be administered in a clinic relatively briefly (≈ 5 min/frequency), and requires little training on the part of the tester. For these reasons, this task has been used in a number of loudness studies (Allen, Hall, & Jeng, 1990; Al-Salim et al., 2010; Brand & Hohmann, 2002; Elberling, 1999; Heeren et al., 2013; Ricketts & Bentler, 1996).

A listener's CLS function, generally computed from the median stimulus level for each response category, is the standard for analysis of the results. The standard approach treats the variability in the response data as something to be removed. We propose to instead model the probabilistic nature of listener responses and use these probabilistic representations to improve loudness measurements and further understand categorical perception. At each stimulus level, there is a probability distribution across loudness categories for each listener; when plotted as a

function of stimulus level, these distributions form a multi-category psychometric function (MCPF). The MCPF is a representation that fully models the probabilistic nature of listeners' categorical perception. This model can be applied to listener simulations or when incorporating probabilistic analysis techniques, such as estimation, information, or detection theory.

We describe such an application of estimation theory to improve the accuracy of the CLS function. The ISO recommendations (Kinkel, 2007) for adaptive CLS testing are designed to constrain the test time to a duration that is clinically acceptable. Due to the relatively low number of experimental trials, the natural variability of responses can create an inaccurate CLS function. Our proposed maximum-likelihood (ML) estimation approach is inspired by the work of Green (1993), who developed a technique for using estimation theory to determine the psychometric function of a "yes-no" task. His approach, when applied to measurements of audiometric threshold, increased the accuracy of the measurements (e.g., Florentine, Buus, & Geng, 2000; Amitay, Irwin, & Moore, 2006). In this paper, we modify and extend this concept to estimate the MCPF that best describes an individual listener's categorical loudness perception.

In this paper, we (1) introduce the concept of the MCPF representation, (2) describe a parameterization method for the representation, (3) use principal components analysis (PCA) to create a representative catalog, (4) show MCPFs for NH and HL listeners, and (5) demonstrate how the catalog can be paired with a ML procedure to estimate an individual's MCPF.

2 Methods

2.1 Multi-Category Psychometric Function

A standard psychometric function describes the probability of a particular response as a function of an experimental variable. We introduce the concept of a multi-category psychometric function (MCPF), which represents the probability distribution across multiple response categories as a function of an experimental variable. A MCPF consists of a family of curves that demarcate the probabilities of multiple categories as a function of stimulus level.

As an example, we demonstrate the construction a hypothetical CLS MCPF for a NH listener, for an 11-category CLS scale. Figure 1 (a) shows the listener's probability distribution across categories at 60 dB SPL. One can see that the loudness judgments are not constrained to one category, but, instead, form a unimodal distribution. Figure 1 (b) shows the cumulative probability density function for the data in (a). The MCPF (Fig. 1 (c)) is constructed by plotting each cumulative distribution (marked by circles in (b)) as a function of stimulus level. In Fig. 1 (c), the probabilities of each categorical response at 60 dB SPL are marked with arrows; one can see that the vertical distance between the curves matches the probabilities shown in Fig. 1 (a,b). The MCPF shows how the probability distribution across loudness-category judgments varies with stimulus level.

2.2 Parameterization

The MCPF curves are logistic functions that can be parameterized. The four-parameter logistic function $f(x|\theta)$, was selected to represent each of the curves of the MCPF. For our 11-category CLS scale, each MCPF, $F(x|\theta)$, consists of 10 category-boundary curves. Thus, the modeling implementation has 10 sets of parameters θ_i that define the 10 curves of each MCPF.

$$f(x|\theta) = D + \frac{A-D}{1+(x/C)^B} \qquad (1)$$

$$\theta = \{A, B, C, D\}$$

$$A = minimum\ asymptote,\ B = slope,\ C = inflection\ point,\ D = maximum\ asymptote$$

$$F(x|\theta) = f_i(x|\theta_i) = D_i + \frac{A_i-D_i}{1+(x/C_i)^{B_i}} \qquad i = 1, \ldots, 10 \qquad (2)$$

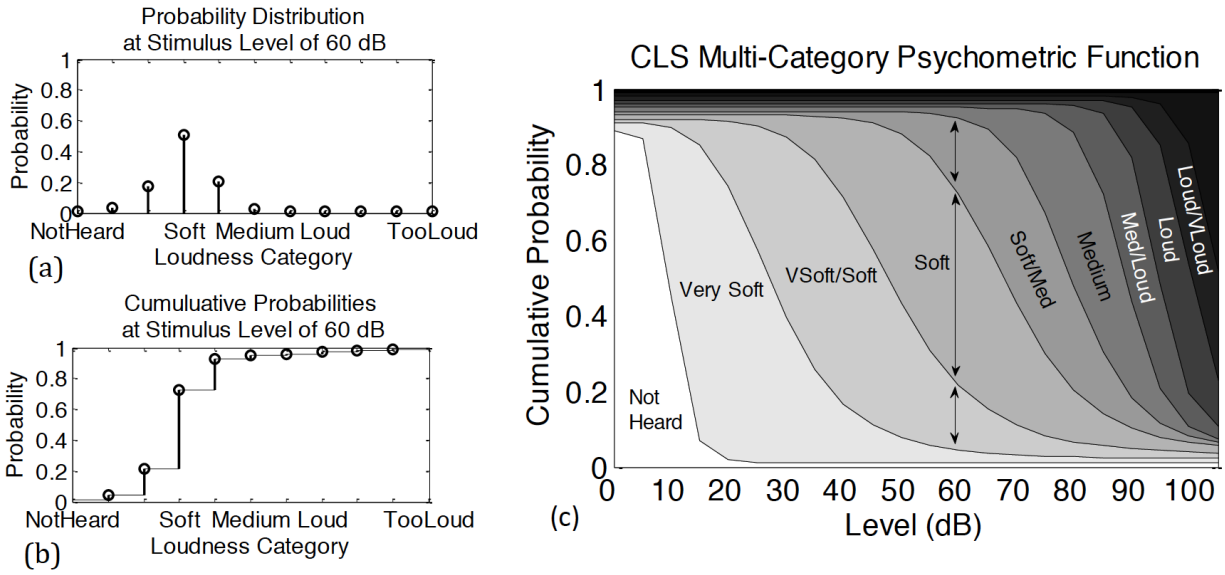


Figure 1. Example showing how the MCPF is constructed from categorical probabilities. (a) Probability distribution across loudness categories for the fixed stimulus level of 60 dB SPL. (b) Cumulative probability function for loudness categories at 60 dB SPL; probabilities are the same as shown in (a). (c) MCPF from 0 to 105 dB SPL. The vertical distance between curves represents the probability of each category. The 3 highest category probabilities at the 60 dB SPL stimulus level are marked by arrows, and correspond to the data in (a, b). Maximum ("Very Loud" and "Too Loud") categories are not labelled.

2.3 A Representative Catalog

We have developed a "catalog" of MCPFs to be representative of both NH and HL listener CLS outcomes. Each function in the catalog has the same form (Eq. 2) and is defined by a set of parameters, $(\theta_i, i = 1, \dots, 10)$; these parameters are varied within the catalog to capture the natural variability across listeners. A PCA analysis decomposes the primary sources of variability in the data. Let the matrix of all listener parameters be $X = \{\theta^1, \dots, \theta^M\}$, where M is the number of listeners, then the set of listener weightings, $W = \{w^1, \dots, w^M\}$, from the projections on the PCA eigenvectors v is $Xv = W$.

The weightings used to construct the parameters for the catalog of functions were determined by evenly sampling from the range of listener weightings, with the number of samples proportionally allocated based on energy. Combinatorial combinations of the sampled weightings were used to construct the catalog. The superset of derived MCPF parameter sets that defines the catalog, is denoted as Θ.

2.4 Maximum-Likelihood Estimation

The catalog can be used to compute a maximum-likelihood (ML) estimate of a listener's MCPF. This may be particularly useful when the number of experimental observations is relatively low, but an accurate estimate of a listener's loudness perception is needed, as is the case in most adaptive-level CLS methods.

We denote a listener's raw CLS data as $(x_1, \dots, x_N)$, where N is the total number of experimental observations. The likelihood, $\mathcal{L}(\cdot)$, of these observations is computed for each catalog MCPF, $F(x|\theta)$, maximizing over all potential parameter sets (i.e., all $\theta \in \Theta$). The maximization is computed over the functionally equivalent log-likelihood.

$$\text{argmax}_{\theta\in\Theta} \ln \mathcal{L}(\theta; x_1, \dots, x_N) = \text{argmax}_{\theta\in\Theta} \sum_{i=1}^{N} \ln F(x_i|\theta) \qquad (3)$$

Once the ML parameter set has been found, it can be used to construct the listener's MCPF, $F(x|\theta)$, via Eqs. 1 and 2.

2.5 Experiment

Participants

Sixteen NH listeners and 25 listeners with HL participated. One ear was tested per participant. NH participants had audiometric thresholds ≤ 10 dB HL; participants with sensorineural HL had audiometric thresholds 15-70 dB HL.

Table 1. NH and HL listener characteristics. N is the number of participants. Age, audiometric threshold, and loudness discomfort level (LDL) are reported. Standard deviations are shown in ( ). Listeners that did not report an LDL from 0-105 dB SPL are tallied as NA.

	N	Age (yr)	Threshold, 1 kHz	Threshold, 4 kHz	LDL, 1 kHz (dB SPL)	LDL, 4 kHz (dB SPL)
NH	15	38 (±10)	2 dB HL (±5)	5 dB HL (±5)	102 (±4) [6 NA]	102 (±4) [9 NA]
HL	22	53 (±16)	31 dB HL (±18)	51 dB HL (±12)	102 (±6) [3 NA]	100 (±4) [11 NA]

Stimuli

Pure-tone stimuli (1000-ms duration, 25-ms onset/off cosine-squared ramps) were sampled at 44100 Hz. CLS was measured at 1 and 4 kHz. Stimuli were generated and presented using MATLAB.

Fixed-Level Procedure

The CLS experimental methods generally followed the ISO standard (Kinkel, 2007). Stimuli were presented monaurally over Sennheiser Professional HD 25-II headphones in a sound booth. Eleven loudness categories were displayed on a computer monitor as horizontal bars increasing in length from bottom to top. Every other bar between the "Not Heard" and "Too Loud" categories had a descriptive label (i.e., "Soft", "Medium", etc.). Participants clicked on the bar that corresponded to their loudness perception.

The CLS test was divided into two phases: (1) estimation of the dynamic range, and (2) the main experiment. In the main experiment, for each frequency, a fixed set of stimuli was composed of 20 repetitions at each level, with the presentation levels spanning the listener's full dynamic range in 5 dB steps. Stimulus order was pseudorandomized such that there were no consecutive identical stimuli and level differences between consecutive stimuli did not exceed 45 dB.

3 Results

3.1 Individual Listener MCPFs

In Fig. 2, we examine how the MCPF extends the representation of CLS data beyond a standard CLS function, which is typically determined from the median level assigned to each loudness category. In each MCPF, note that the vertical distance between the upper and lower boundary curves for a category is the probability of that category, and the horizontal distance between boundary curve intercepts can be thought of as the "width" of a category, in dB. The steeper the slope of a curve, the more defined the distinction between categories, whereas shallower curves coincide with probability being distributed across two or more categories over a range of levels. A wider probability distribution across categories, at a fixed level, indicates that the listener had more uncertainty (i.e. entropy) in their responses.

The top two rows of Fig. 2 show data for two representative NH listeners, NH04 at 1 kHz (1st row) and NH01 at 4 kHz (2nd row). Both CLS functions show the characteristic loudness growth that has been documented for NH listeners in the literature. For listener NH04, a single category

dominates the response probability ($P \geq 0.8$) at the lowest stimulus levels. Above 25 dB SPL, although the probability becomes more distributed across categories, no more than 3 categories show greater than 10% response probability at any fixed stimulus level, indicating that the listener had a low amount of uncertainty in their categorical perception. Compared to NH04, listener NH01 has a wider range of levels that correspond to 5 CU ("Very Soft") and a lower LDL, leaving a smaller dynamic range for the remaining categories. This results in a horizontally-compressed MCPF, mirroring the listener's sharper growth of loudness in the CLS function. The category boundary curves for this listener are also shallower from 30-80 dB SPL, corresponding to probability distributions with higher variance. The bottom two rows of Fig. 2 show representative data for listeners with varying degrees of HL. Listener HL17 (3rd row) shows roughly even widths of each loudness category ($\approx 10\ dB\ SPL$). Data for HL11 (4th row) is an example of a listener with hearing loss and an LDL that is within the ranges of a NH listener; the higher threshold level causes the spacing between category boundary curves to be compressed horizontally. Despite this compression and hearing loss, this listener with HL has well-defined perceptual boundaries between categories, observable from the sharp slopes of the category boundary curves.

These NH and HL examples show some general patterns across both the 1 and 4 kHz stimuli. We see that the steepness of the MCPF slopes reflects the uncertainty of a listener's responses, and that the most uncertainty is observed across the 5 CU to 25 CU categories. The most across-listener variability in category width is observed for 5 CU, which spans in width from a maximum of 50 dB (NH06) to a minimum of 3 dB (HL06). A higher threshold and/or lower LDL results in a horizontally compressed MCPF; this compression narrows all intermediate categories.

3.2 Construction of the MCPF Catalog

A PCA of the combined NH and HL data revealed that the first two eigenvectors were sufficient for capturing at least 90% of the variability in listener parameters. Vector weightings for creating the MCPF catalog were evenly sampled from the range (2 standard deviations) of the individual listener weightings. Sixty-six sampled weightings were combined to create the 1460 MCPFs that constitute the catalog

3.3 Application to ML estimation

The MCPF catalog contains models of loudness perception for a wide array of listener types. This probabilistic catalog can be used in a variety of applications. Here, we demonstrate how a ML technique can be used with the catalog to estimate a novel listener's MCPF from a low number (≤ 30) of CLS experimental trials.

Five additional NH listeners, whose data were not used in the construction of the catalog, were recruited to complete a CLS task. In addition to the set of tasks described in the Methods section, these listeners completed a shorter-duration CLS task that used an adaptive stimulus-level selection technique that conformed to the ISO standard. As opposed to the fixed-level task, which presented stimuli at each level 20 times and spanned the dynamic range in 5-dB intervals, the adaptive technique selected 10 levels that spanned the dynamic range and presented these 3 times. For reference, for a listener with a 0-105 dB SPL dynamic range, the fixed-level technique required 440 experimental trials while the adaptive-level technique required 30 trials (i.e., reduced by more than an order of magnitude).

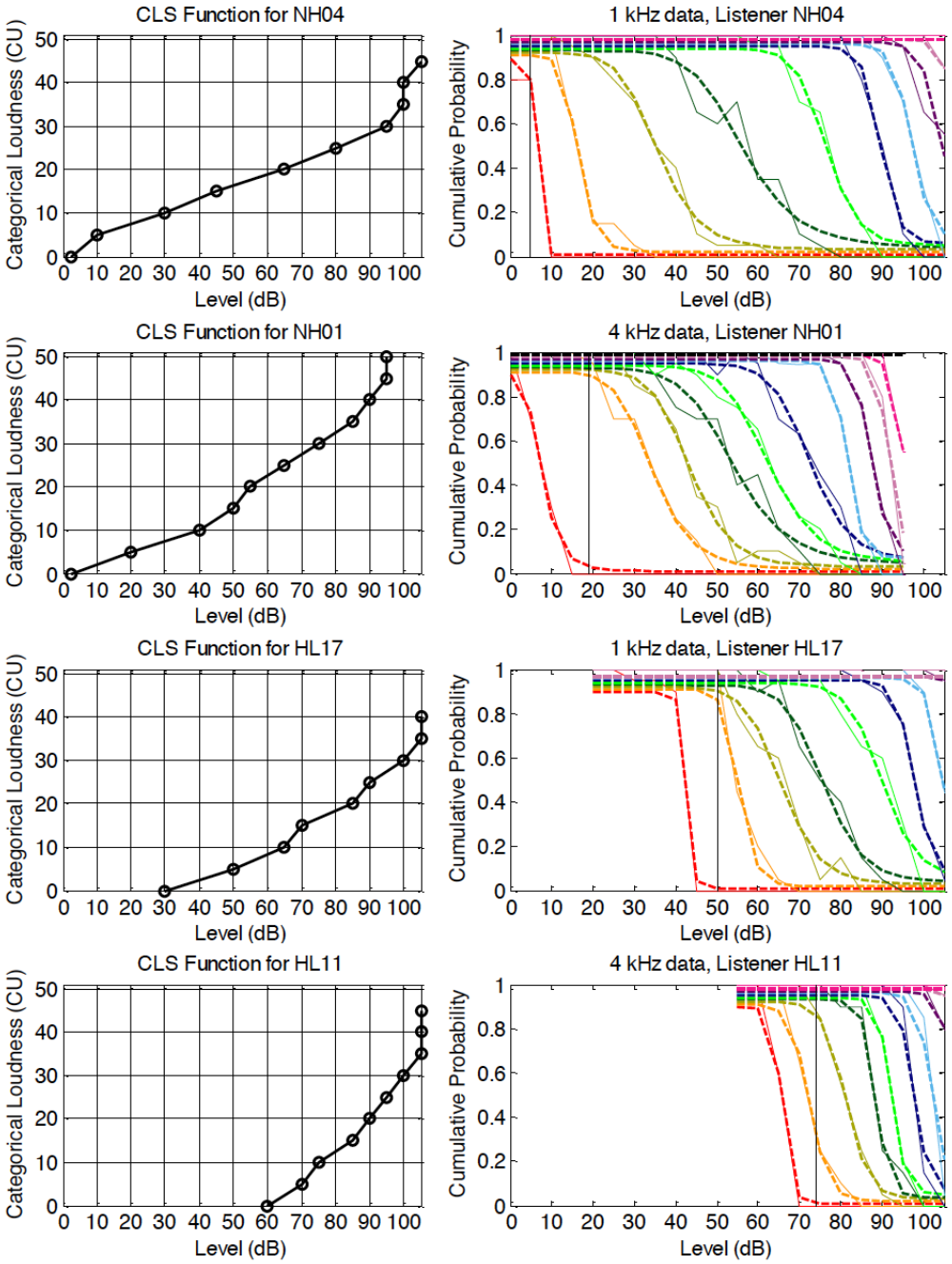


Figure 2. The CLS function and corresponding MCPF for 4 listeners: (1st row) listener NH04 for 1 kHz stimuli, (2nd row) listener NH01 for 4 kHz stimuli, (3rd row) listener HL17 for 1 kHz stimuli, (4th row) listener HL11 for 4 kHz stimuli. The CLS function plots the median level for each CU. The MCPFs show the raw data as solid lines and the parameterized logistic fits as dashed lines. Each listener's audiometric threshold is marked with a vertical black line.

The result of the ML estimation is a MCPF that fully describes the probabilistic behavior of the listener. As the adaptive technique has a maximum of 3 presentations at each stimulus level, a histogram-estimation of the MCPF from this sparse data is probably inaccurate. In addition to computing the most likely MCPF, the 50% intercepts of the ML category boundary curves may be used to estimate the CLS function, as shown in Fig. 3. To compute the average root mean squared error (RMSE) across these 5 NH listeners and two frequencies, we use each listener's results from fixed stimulus-level experiment as the 'true' CLS function values. The average RMSE for the median of the adaptive stimulus-level technique was 6.1 dB, and the average RMSE for the ML estimate, based on the same adaptive technique data, was 4.4 dB. The ML estimate

improves the accuracy of the CLS function and provides a probabilistic model of the listener's loudness perception.

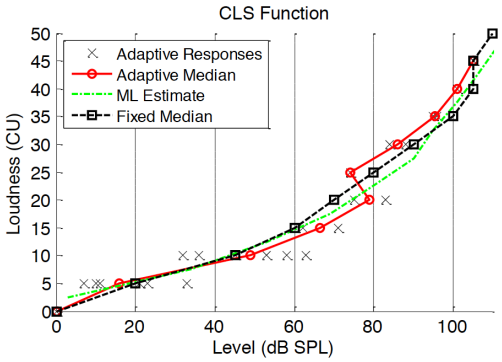


Figure 3. CLS function for a NH listener. Raw data from the ISO adaptive data collection are shown with x's. The median level for each CU is shown with red circles. The resulting median CLS function based on the long-term fixed-stimulus experiment is shown with black squares, representing the 'best' estimate of the listener's loudness perception. The ML-based estimate is shown as a green dashed line.

4 Discussion

The CLS MCPF describes how categorical perception changes with stimulus level. The results demonstrate that loudness perception of NH and HL listeners is a random variable at each stimulus level, i.e., each level has a statistical distribution across loudness categories.

The development of a probabilistic model for listener loudness perception has a variety of advantages. The most common usage for probabilistic listener models is to simulate listener behavior for the development of experiments or listening devices. Probabilistic models also allow one to apply concepts from detection, information, and estimation theory to the analysis of results and the methodology of the experiment. In this paper, we show how a catalog of MCPFs can be used to find the ML estimate of a listener's MCPF, when a relatively small number of experimental trials are available. The catalog of MCPFs can also be used to develop an adaptive stimulus-selection technique that minimizes the entropy of the resulting outcome with each experimental trial. These MCPF representations extend our knowledge of the variability of the listener decision process.

The ISO adaptive recommendation for CLS testing results in a relatively low number of experimental trials ($\approx$ 15 to 30), in order reduce the testing time and make the test more clinically realizible. Although this is an efficient approach, due to the low number of samples, the resulting CLS function can be unreliable. The ML estimate of a listener's loudness perception based on this lower number of experimental trials is able to accurately predict a listener's long-term CLS function, without removing outliers or using assumptions about the shape of the function to smooth the result. The MCPF may be further exploited to select experimental stimulus levels adaptively, such that each presentation maximizes the expected information. The ML MCPF estimate provides greater insight into the nature of the listener's categorical perception (Torgerson, 1958) and the potential for optimal measurement methods, while still allowing for clinically-acceptable test times.

Acknowledgements

This research was supported by the NIH grants: T32 DC000013, R01 DC011806 (WJ), R01 DC008318 (STN), and P30 DC004662.

References

Allen, J.B. (2008). Nonlinear cochlear signal processing and masking in speech perception. Springer Handbook of Speech Processing, 27-60.

Allen, J.B., Hall, J.L. & Jeng, P.S. (1990). Loudness growth in 1/2-octave bands (LGOB)—A procedure for the assessment of loudness. JASA 88(2), 745-753.

Al-Salim, S.C., Kopun, J.G., Neely, S.T., Jesteadt, W., Stiegemann, B. & Gorga, M.P. (2010). Reliability of categorical loudness scaling and its relation to threshold. Ear and Hearing 31(4), 567-578.

Amitay, S., et al. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. JASA 119, 1616-1625.

Brand, T. & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. JASA 112(4), 1597-1604.

Elberling, C. (1999). Loudness scaling revisited. JAAA 10(5), 248-260.

Florentine, M. (2011). Loudness. Springer New York.

Florentine, M., Buus, S. & Geng, W. (2000). Toward a clinical procedure for narrowband gap detection: I. A psychophysical procedure. Audiology 39, 161-167.

Green, D.M. (1993). A maximum-likelihood method for estimating thresholds in a yes–no task. JASA 93(4), 2096-2105.

Heeren, W., Hohmann, V., Appell, J. E., & Verhey, J. L. (2013). Relation between loudness in categorical units and loudness in phons and sones. JASA, 133(4), EL314-EL319.

Kinkel, M. (2007). The new ISO 16832 'Acoustics–loudness scaling by means of categories'. 8th EFAS Congress/10th Congress of the German Society of Audiology; Heidelberg.

Ricketts, T.A. & Bentler, R.A. (1996). The effect of test signal type and bandwidth on the categorical scaling of loudness. JASA 99(4), 2281-2287.

Torgerson, W. (1958). Theory and methods of scaling. New York: Wiley.

Chapter 18

Auditory fMRI of Sound Intensity and Loudness for Unilateral Stimulation

Oliver Behler[1], Stefan Uppenkamp[1,2]

[1]Medizinische Physik and [2]Cluster of Excellence Hearing4All, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Contact author: OB, oliver.behler@uni-oldenburg.de

Abstract

We report a systematic exploration of the interrelation of sound intensity, ear of entry, individual loudness judgments, and brain activity across hemispheres as well as a functional differentiation of regions within auditory cortices using auditory functional magnetic resonance imaging (fMRI). The stimuli employed were 4-kHz-bandpass filtered noise stimuli, presented monaurally to each ear at levels from 37 to 97 dB SPL. One diotic condition and a silence condition were included as control conditions. Normal hearing listeners completed a categorical loudness scaling procedure prior to listening to similar stimuli while auditory fMRI was performed. The relationship between brain activity, as inferred from blood oxygenation level dependent (BOLD) contrasts, and both sound intensity and loudness estimates were analyzed by means of functional activation maps and linear mixed effects models for various anatomically defined regions of interest in the ascending auditory pathway and in the cortex. The results indicate distinct functional differences between midbrain and cortical areas as well as between specific regions within auditory cortex, suggesting a systematic hierarchy in terms of lateralization and the representation of sensory stimulation and perception.

Keywords: neural activation, sound pressure level, categorical loudness scaling, auditory cortex, auditory pathway, monaural stimuli

1 Introduction

Loudness is mainly the perceptual correlate of sound intensity, which is usually expressed in dB SPL as sound pressure level on a logarithmic scale. However, loudness judgments are also affected by several other variables, including physical sound parameters like e.g. duration, spectral content and temporal modulation as well many more acoustic and non-acoustic factors, including the particular task or context and personal factors like the individual hearing status (Florentine et al., 2011). While the relationship between sound intensity and neural activity in the human central auditory pathway has been extensively studied by means of neuroimaging techniques, only a small number of studies have investigated the interrelation of sound intensity, loudness and the corresponding brain activity (for a review, see Uppenkamp & Röhl, 2014). Some auditory functional magnetic resonance imaging (fMRI) studies indicate that neural activation, at least in auditory cortex (AC), might be more a representation of perceived loudness rather than of physical sound pressure level (Hall et al., 2001; Langers et al., 2007; Röhl & Uppenkamp, 2012). The current literature still does not provide definite answers to the following questions: (1) At what stage or stages along the auditory pathway is sound intensity

transformed into its perceptual correlate (i.e. loudness)? (2) What are the functional differences across regions within AC with respect to loudness-related activation? Promoting a better understanding of the neural basis of loudness might, in the long run, improve diagnostics and treatment of hearing disorders characterized by a distorted loudness perception, e.g. loudness recruitment.

The present study is aimed at extending the current literature by providing a detailed characterization of the neural representation of sound intensity and loudness, as reflected by functional MRI. In a group of normal hearing listeners, we systematically explored the interrelation of ear of entry, sound pressure level, individual loudness and brain activation, as defined by the blood oxygenation level dependent (BOLD) signal, in the ascending auditory pathway and within AC.

2 Methods

2.1 Participants and Procedure

Thirteen normal hearing volunteers (aged 34 ± 8 years, 4 females) participated in this study. Each participant attended two experimental sessions. In the first session, standard audiometry and an adaptive categorical loudness scaling procedure (Brand & Hohmann, 2002) were performed in a sound booth. In the second session, auditory fMRI was performed while subjects were doing a simple listening task in the MRI scanner.

2.2 Stimuli

All stimuli consisted of 1/3 octave band-pass low-noise noise (Pumplin, 1985) bursts at 4 kHz center frequency and were delivered via MRI compatible insert earphones (Sensimetrics S14, Sensimetrics Corporation, Malden, MA). In the loudness scaling procedure, single noise bursts with a maximum intensity of 105 dB SPL were used under left monaural, right monaural and diotic stimulus conditions. In the MRI experiment, trains of noise bursts with a total duration of 4.75 s were presented left and right monaurally at 37, 52, 67, 82 and 97 dB SPL and diotically at 82 dB SPL.

2.3 MRI data acquisition

Functional and structural images were acquired with a 3-Tesla MRI system (Siemens MAGNETOM Verio). Functional images were obtained using T2*-weighted gradient echo planar imaging (EPI), with a sparse temporal sampling paradigm to reduce the influence of the acoustic noise created by the scanner (Hall et al., 1999). Stimuli were presented in pseudorandomized order during 5-s gaps of scanner silence in between two successive volume acquisitions. Stimuli of each of the eleven conditions plus a silence condition, which served as baseline, were presented 36 times over the course of the experiment. For the purpose of maintaining the participants' attention towards the acoustic stimuli, they were asked to count the number of occasionally presented deviants, characterized by a transient dip in sound level in one of the noise bursts.

2.4 Psychoacoustic evaluation

Individual loudness judgments obtained in the scaling procedure were used to fit loudness functions for each participant by means of a recently suggested fitting method (Oetting et al., 2014). Loudness estimates for the stimulus conditions used in the MRI experiment were extracted from the individual loudness functions and were used for further analyses.

2.5 MRI data analysis

Standard preprocessing of the imaging data (including spatial smoothing with a 5-mm FWHM Gaussian kernel) and general linear model (GLM) estimation was done using SPM8 (Functional Imaging Laboratory, The Wellcome Department of Imaging Neuroscience, London, UK, http://www.fil.ion.ucl.ac.uk/spm/). Three different general linear models were set up to model the BOLD signal in every voxel as a function of ear of entry and either sound pressure level or loudness estimates for each participant. In each model, the silence condition was implicitly modeled as baseline. Model 1 included one separate regressor for each of the remaining 11 stimulus conditions. Model 2 and 3 contained one regressor for the binaural condition and two categorical regressors for left and right monaural stimuli, extended by 1st and 2nd order parametric regressors (the 2nd orthogonalized to the 1st) for sound intensity (model 2) or categorical loudness (model 3). Based on the resulting individual (1st level) coefficient maps, two approaches were carried out to analyze the relationship between neural activation and sound intensity or loudness for left or right stimuli across subjects (at the 2nd level):

(1) Functional activation map trend analysis: The 1st level contrast images for the parametric regressors estimated in model 2 and 3 were entered into separate one sample t-tests. The resulting 2nd level t-statistic parametric maps were used to detect voxels characterized by a significant positive linear or quadratic relationship (over and above the linear trend) of measured signal with sound intensity or individual loudness for left or right ear stimulation. For the purpose of descriptive anatomical localization, statistical parametric maps were overlaid onto the group mean structural image, using the MRIcron software (version 6, 2013; Chris Rorden, Columbia, SC; http://www.mccauslandcenter.sc.edu/mricro/mricron/).

(2) Region of interest (ROI) analysis: For each participant, twelve auditory ROIs were defined based on anatomical landmarks in the individual structural images: Left and right inferior colliculus (IC), medial geniculate body (MGB), Planum temporale (PT), posterior medial (HGpm), central (HGc) and anterolateral (HGal) parts of the first Heschl's gyrus. The average signal change from baseline for every stimulus condition of all voxels within spheres of 5-mm radius centered at individual ROI coordinates was calculated based on the regression coefficients of model 1 and entered into random slope linear mixed effects models (LMMs). For each of the twelve ROIs, eight separate LMMs (2 x 2 x 2) were estimated modeling the ROI percent signal change as a linear or a quadratic function of sound intensity or individual loudness, for left and right ear of entry. Model parameters were estimated by means of maximum-likelihood. Likelihood-ratio tests were conducted to assess significance of the models: First, linear models were tested against "null models" containing only the constant terms. Then, quadratic models were tested against the corresponding linear models. In both steps, models were considered significant at a level of $p < .05$, Bonferroni-corrected. To provide measures of the models' goodness-of-fits in terms of explanatory power, marginal R^2 statistics (R^2_m), representing the partition of variance in the model explained by the fixed (population) effects, were calculated as suggested by (Johnson, 2014).

3 Results

3.1 Categorical Loudness Scaling

Figure 1 shows the results of the categorical loudness scaling procedure for the group of 13 participants. Group averaged fitted loudness curves (Oetting et al., 2014) for left monaural, right monaural and binaural stimuli are shown, along with the interindividual

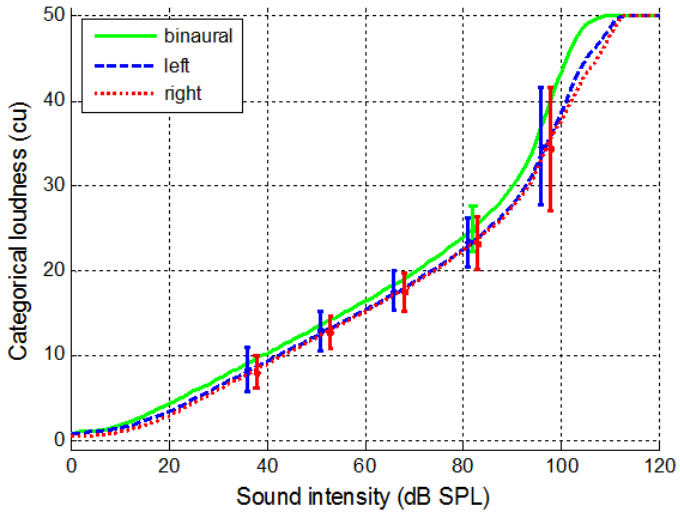


Figure 1: Categorical loudness as a function of sound intensity and ear of entry. The three curves represent group averages of individual loudness fits. Error bars represent interindividual standard deviations of loudness estimates for the stimulus intensities presented in the MRI experiment.

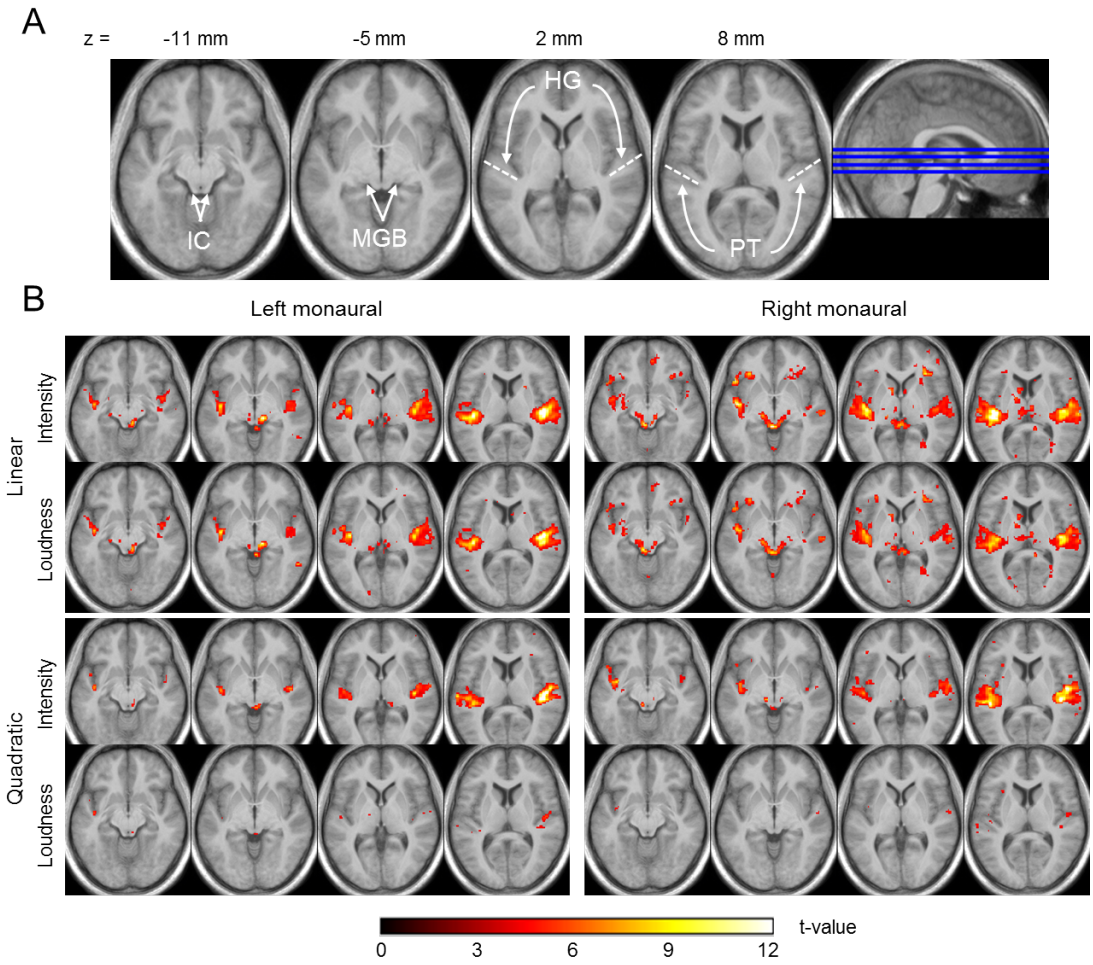


Figure 2: Trend analysis for brain activation as a function of sound intensity and categorical loudness for left and right monaural stimuli. Second-level t-statistic parametric maps for the contrasts of the linear and quadratic (orthogonal) regressors are thresholded at a significance level of $p < 0.001$ (uncorrected, $t > 3.93$) and overlaid onto the group mean structural image at four axial slices corresponding roughly to the group average anatomical locations of the auditory ROIs: z = -11 → IC; z = -5 → MGB; z = 2 → central and anterolateral HG; z = 8 → posterior medial HG and PT.

standard deviations of loudness estimates for the stimulus intensities presented in the MRI experiment. All three curves are characterized by a nearly linear growth of categorical loudness with sound intensity between 20 and 80 dB SPL and an increase in the steepness of the slope around 90 dB SPL. There was virtually no difference in perceived loudness between left and right ear of entry. For diotic stimulation, the expected 3 dB effect of binaural loudness summation is clearly visible.

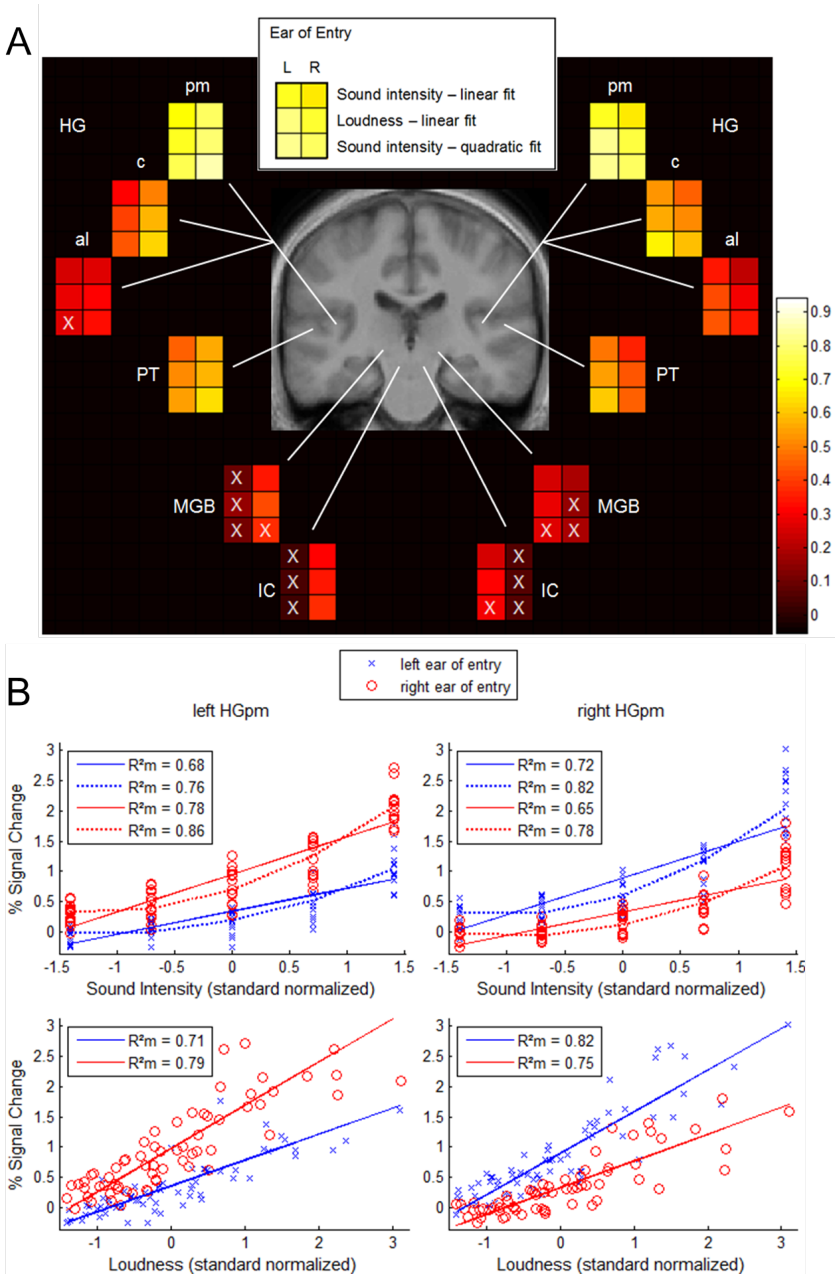


Figure 3: Results of the ROI analysis. **A:** Marginal R^2 (R^2_m) statistics, representing the partition of variance explained by fixed effects, of linear and quadratic model fits with sound intensity as well as linear fits with loudness for left (L) and right (R) monaural stimuli corresponding to each auditory ROI. Non-significant models are marked with a white "X". R^2_m values for quadratic fits with loudness are not shown, since the corresponding models reached significance in only 4 out of 24 ROIs. **B:** Percent signal change from baseline in the left and right HGpm plotted as a function of ear of entry and sound intensity or categorical loudness. Overlaid are the fitted curves as predicted by the fixed effects of the linear (straight lines) and quadratic (dotted curves) models with their corresponding R^2_m statistics.

3.2 Functional Activation Map Trend Analysis

Figure 2 illustrates the results of the functional activation map trend analysis probing linear and quadratic growths of activation magnitude with sound intensity and loudness for left and right ear of entry. The activation patterns for the contrasts of the linear intensity and loudness regressors appear highly similar. Significant clusters of voxels were found in all sections of HG, in PT, MGB and IC in both hemispheres for right monaural stimuli and in all of these regions except for the left IC and HGal for left monaural stimuli. For both monaural conditions, the extent of activation was generally more pronounced in the respective contralateral hemisphere. In contrast, the analysis revealed considerable differences between the contrasts of the quadratic regressors: While large clusters of voxels in AC and smaller ones in IC and MGB were characterized by a significant quadratic signal growth with sound intensity over and above the linear trend, only very few voxels displayed a quadratic relationship with categorical loudness.

3.3 Region-of-Interest Analysis

The results of the ROI analysis are illustrated in Figure 3. R^2_m statistics, representing that part of variance explained by the fixed (population) effects, display a largely symmetrical pattern across hemispheres. The highest values were found in the posterior medial parts of Heschl's gyri, whereas the subcortical ROIs and left HGal were characterized by the lowest R^2m values. Throughout all investigated regions, sound levels or loudness estimates provided a higher explanatory power for contralateral as compared to ipsilateral stimuli, albeit with varying degrees of lateralization across ROIs. Linear loudness models, although being outmatched by quadratic fits with sound intensity in the majority of cortical ROIs, still yielded slightly better goodness-of-fits as compared to linear sound intensity models in all regions. According to LLR tests, quadratic loudness models reached significance in only 4 out of 24 ROIs, indicating that the addition of a second polynomial degree did not significantly improve the likelihood of loudness models in the majority of ROIs. The example plots of the different model fits with sound intensity and individual loudness estimates in the left and right HGpm reveal that the significant quadratic component with respect to the signal growth as a function of sound intensity is attributable to a steeper increase of BOLD responses at the highest sound levels (in this regard, the examples are representative for all auditory regions).

4 Discussion

We used auditory fMRI to measure responses to unilateral narrowband noise stimuli in the human auditory system of normal hearing subjects. Based on these measures, we investigated the interrelation of ear of entry, sound pressure level, individual loudness and the corresponding brain activity by means of functional activation maps as well as linear mixed effects models for a large number of distinct regions of interest in the ascending auditory pathway and in the cortex. This approach allowed us to characterize the neural representation of sound intensity and loudness in a detailed way.

4.1 Response characteristics in relation to sound intensity

Throughout all investigated stages of the auditory system, except for ipsilateral stimuli in IC bilaterally and in left MGB, neural activation as reflected by the fMRI BOLD response was significantly and positively related to physical sound intensity. Specifically, at cortical level, changes in sound pressure level were reflected by nonlinear (quadratic) increases of activation magnitude, with steeper slopes at the highest sound levels. While this is in agreement with some previous findings (Hart et al., 2002, 2003), other studies reported a more or less linear increase of BOLD signal strength with sound intensity (Hall et al., 2001; Langers et al., 2007; Röhl & Uppenkamp, 2012) and still others found indications of response saturation at the highest levels (Mohr et al., 1999). To what extent these differences can be explained in terms of the use of

different types of stimuli, dynamic ranges, fMRI paradigms or other factors (e.g., perceived loudness) is an interesting topic by itself, but this is not the scope of the present study. As opposed to AC, activation in the investigated subcortical structures increased predominantly linearly with increasing sound pressure level, which is in line with previous findings (Röhl & Uppenkamp, 2012).

Our findings further suggest considerable differences along the auditory pathway and across auditory regions in cortex with respect to the strength of the relationship between neural activity and sound intensity. The relationship was most pronounced in in the posterior medial section of HG (which includes primary AC), a little less so in PT and comparatively weak in the anterolateral sections of HG and in subcortical regions. These findings largely conform to earlier studies reporting the most robust level dependence in primary AC, but less consistent results with regard to PT and a generally weak sensitivity to level changes in more lateral and anterior regions of HG (Hart et al., 2002; Brechmann et al., 2002; Ernst et al., 2008).

4.2 Interrelation with the ear of entry

Activity at all investigated stages of the auditory system appeared to be more closely related to changes in sound level for stimuli presented at the contralateral ear. This is in good agreement with the anatomy and with previous investigations suggesting that auditory information crosses to the contralateral hemisphere in the lower brain stem followed by ipsilateral signal propagation towards AC (e.g., Langers et al., 2005). However, in the present study, the degree of lateralization differed considerably between regions, being particularly pronounced in the IC of the brain stem and rather weak in the posteromedial and anterolateral sections of HG. In fact, in stark contrast to the IC, changes in sound intensity were highly predictive of signal changes in HGpm irrespective of the ear of entry, pointing towards strong interhemispheric connections in the upper stages of the auditory system.

4.3 Transformation of sound intensity into perceived loudness

As expected from the strong relationship between sound intensity and loudness, a very similar pattern of differences between auditory regions was found for the relationship between BOLD responses and individual loudness. Nonetheless, it was possible to identify distinct differences with regard to the representation of sensory stimulation and perception in the human auditory system. Specifically, while changes in sound pressure level were reflected by a quadratic growth of activation magnitude, especially in cortical areas, the relation between activation and categorical loudness can be described as predominantly linear in all investigated auditory regions. This finding, in conjunction with the similarities between the growth of categorical loudness (Fig. 1) and BOLD signal changes (Fig. 3B) with sound intensity, lends support to previous auditory fMRI studies suggesting that neural activation in AC might be more a direct linear reflection of perceived loudness rather than of physical sound pressure level (Langers et al., 2007; Röhl & Uppenkamp, 2012). By contrast, in the previous study by Röhl & Uppenkamp (2012) no significant relationship between the BOLD signal strength and categorical loudness was found at the level of the IC or MGB. One interpretation in that study was that at the level of the IC, “neural activation might be more a reflection of physical sound intensity”. The current results are different, given that activity in subcortical regions was - albeit only slightly - more closely related to individual loudness as compared to sound pressure. However, the interpretation of BOLD fMRI results from these regions appears difficult, given the poor signal-to-noise ratio caused by the small size of the structures and their susceptibility to cardiac-related, pulsatile motion effects. Taking measures to improve the signal quality in subcortical regions, such as cardiac gated image acquisition (Guimaraes et al., 1998), is therefore strongly advised for future investigations.

The largely linear relationship between categorical loudness and sound intensity over the range of presented levels, as well as the fact that our participants were from a rather homogenous group of normal hearing listeners with similar loudness perception, might have limited the

ability to differentiate between the contributions of both measures to neural activation in the human auditory system. A promising next step would be the inclusion of hearing impaired participants characterized by a more diverging relationship between sound intensity and perceived loudness (e.g., loudness recruitment). Moreover, this approach could provide additional valuable information about the nature of hearing disorders associated with a distorted loudness perception. This in the end might open a potential for the diagnostic use of auditory fMRI. Although it is yet to prove how well the present findings pertain to the hearing impaired listeners, our results so far point towards the activation of the posterior medial section of Heschl's gyrus as the most reliable indicator of individually perceived loudness.

4.4 Conclusion

Our findings support the notion that neural activation in auditory cortex as well as in certain stages of the ascending auditory pathway is more a direct linear reflection of perceived loudness rather than of physical sound intensity. There are, however, systematic differences between the investigated auditory regions in terms of the strength of this relationship and the degree of lateralization. We therefore suggest that functional differentiation, both between cortical and subcortical regions as well as between regions of auditory cortex, is an important issue to consider in the pursuit of a complete and comprehensive understanding of the physiological correlates of loudness perception.

Acknowledgments

Oliver Behler is funded by the PhD programme 'Signals and Cognition' (Niedersächsisches Ministerium für Wissenschaft und Kultur).

References

Brand, T., Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, 112, 1597–1604.

Brechmann, A., Baumgart, F., Scheich, H. (2002). Sound level-dependent representation of frequency modulations in human auditory cortex: A low-noise fMRI study. *Journal of Neurophysiology*, 87(1), 423-433.

Ernst, S.M.A., Verhey, J.L., Uppenkamp, S. (2008). Spatial dissociation of changes of level and signal-to-noise ratio in auditory cortex for tones in noise. *Neuroimage,* 43(2), 321-328.

Florentine, M., Popper, A.N., Fay, R.R. (Eds., 2011). *Loudness. Springer Handbook of Auditory Research, vol. 37*. New York: Springer.

Guimaraes, A.R., Melcher, J.R., Baker, J.R., Ledden, P., Rosen, B.R., Kiang, N.Y.S., Fullerton, B.C., Weisskopf, R.M. (1998). Imaging subcortical auditory activity in humans. *Human Brain Mapping*, 6, 33–41.

Hall, D.A., Summerfield, A.Q., Goncalves, M.S., Foster, J.R., Palmer, A.R., Bowtell, R.W. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping*, 7, 213-223.

Hall, D.A., Haggard, M.P., Summerfield, A.Q., Akeroyd, M.A., Palmer, A.R. (2001). Functional magnetic resonance imaging of sound-level encoding in the absence of background scanner noise. *The Journal of the Acoustical Society of America*, 109, 1559-1570.

Hart, H.C., Palmer, A.R., Hall, D.A. (2002). Heschl's gyrus is more sensitive to tone level than non-primary auditory cortex. *Hearing Research*, 171(1-2), 177-190.

Hart, H.C., Hall, D.A., Palmer, A.R. (2003). The sound-level-dependent growth in the extent of fMRI activation in Heschl's gyrus is different for low- and high-frequency tones. *Hearing Research*, 179(1-2), 104-112.

Johnson, P.C.D. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slope models. *Methods in Ecology and Evolution*, 5(9), 944-946.

Langers, D.R.M., van Dijk, P., Backes, W.H. (2005). Lateralization, connectivity and plasticity in the human central auditory system. *Neuroimage*, 28(2), 490-499.

Langers, D.R.M., van Dijk, P., Schoenmaker, E.S., Backes, W.H. (2007). fMRI activation in relation to sound intensity and loudness. *Neuroimage*, 35(2), 709-718.

Mohr, C.M., King, W.M., Freeman, A.J., Briggs, R.W., Leonard, C.M. (1999). Influence of speech stimuli intensity on the activation of auditory cortex investigated with functional magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 105, 2738-2745.

Oetting, D., Brand, T., Ewert, S.D. (2014). Optimized loudness-function estimation for categorical loudness scaling data. *Hearing Research*, 316, 16-27.

Pumplin, J. (1985). Low-noise noise. *The Journal of the Acoustical Society of America*, 78, 100-104.

Röhl, M., Uppenkamp, S. (2012). Neural coding of sound intensity and loudness in the human auditory system. *Journal of the Association for Research in Otolaryngology*, 13(3), 369-379.

Uppenkamp, S., Röhl, M. (2014). Human auditory imaging of sound intensity and loudness. *Hearing Research*, 307, 65-73.

Chapter 19

Tinnitus- and Task-Related Differences in Resting-State Networks

C.P.Lanting, A. Woźniak, P. van Dijk, D.R.M. Langers

University Medical Center Groningen, Department of Otorhinolaryngology, Groningen, Netherlands
University of Groningen, Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, Groningen, Netherlands

Contact author: CPL, c.p.lanting@umcg.nl

Abstract

We investigated tinnitus-related differences in functional networks in adults with tinnitus by means of a functional connectivity study. Previously it was found that various networks show differences in connectivity in patients with tinnitus compared to controls. How this relates to patients' ongoing tinnitus and whether the ecological sensory environment modulates connectivity remains unknown. Twenty healthy controls and twenty patients suffering from chronic tinnitus were enrolled in this study. Except for the presence of tinnitus in the patient group, all subjects were selected to have normal or near-normal hearing. fMRI data were obtained in two different functional states. In one set of runs, subjects freely viewed emotionally salient movie fragments ("fixed-state") while in the other they were not performing any task ("resting-state"). After data pre-processing, Principal Component Analysis was performed to obtain 25 components for all datasets. These were fed into an Independent Component Analysis (ICA), concatenating the data across both groups and both datasets, to obtain eleven group-level networks of neural origin, each consisting of spatial maps with their respective time-courses. Subject-specific maps and their time-course were obtained by back-projection (Dual Regression). For each of the components a mixed-effects linear model was composed with factors group (tinnitus vs. controls), task (fixed-state vs. resting state) and their interaction. The neural components comprised the visual, sensorimotor, auditory, and limbic systems, the default mode, dorsal attention, executive-control, and frontoparietal networks, and the cerebellum. Most notably, the default mode network (DMN) was less extensive and significantly less strongly active in tinnitus patients than in controls. This group difference existed in both paradigms. At the same time, the DMN was stronger during resting-state than during fixed-state in the controls but not the patients. We attribute this pattern to the unremitting engaging effect of the tinnitus percept.

Keywords: Tinnitus, Resting-state fMRI, Independent Component Analysis, functional connectivity

1 Introduction

Tinnitus is a percept of sound that is not related to an acoustic source outside the body. For many forms of tinnitus, mechanisms in the central nervous system are believed to play a role in the pathology. Despite its high prevalence of 5-10%, relatively little is known about the neural mechanisms and causes of tinnitus. Neuroimaging methods have been applied in the last decade to study potential mechanisms, and techniques such as electroencephalography (EEG),

magnetoencephalography (MEG), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) have been extensively used to investigate neural correlates of tinnitus. Previously, neuroimaging studies on tinnitus relied mostly on task-related paradigms, such as presentation of sound stimuli that elicit sound-evoked responses, or manipulation of patients' tinnitus by somatic modulation (i.e., jaw protrusion or a change of gaze) to temporarily affect patients' tinnitus (Lanting, de Kleine, & van Dijk, 2009). Alternatively, studies focused on anatomical differences in brain structure (Adjamian, Hall, Palmer, Allan, & Langers, 2014).

However, since a few years, new data-analysis techniques, such as independent component analysis (ICA), have emerged allowing for identification of coherent patterns of spatially independent, temporally correlated signal fluctuations in neuroimaging data during "resting-state", that is, without a predefined task. These patterns have since been termed "resting-state networks" (RSNs) (M. D. Greicius, Srivastava, Reiss, & Menon, 2004) and represent functionally relevant and connected networks or systems supporting core perceptual and cognitive processes. Resting-state networks are generally reported to show reliable and consistent patterns of functional connectivity (Zhang et al., 2008).

Moreover, advances in data-analysis currently allow for between-group analyses where individual networks can be compared across tasks and groups. Recently, resting-state fMRI has been used to investigate functional connectivity differences in individuals with tinnitus (Burton et al., 2012; Davies, Gander, Andrews, & Hall, 2014; Kim et al., 2012; a Maudoux et al., 2012; A. Maudoux et al., 2012). Results indicate differences in connectivity between healthy and affected subject groups. One result showed reduced connectivity between the left and right auditory cortex, indicative of disturbed excitation and inhibition across hemispheres (Kim et al., 2012). A different study found that in individuals with tinnitus connectivity was increased in numerous brain regions, including the cerebellum, parahippocampal gyrus, and sensorimotor areas. Decreased connectivity was found in the right primary auditory cortex, left prefrontal, left fusiform gyrus, and bilateral occipital regions (a Maudoux et al., 2012; A. Maudoux et al., 2012). Very recently, altered interhemispheric functional connectivity was shown and linked with specific tinnitus characteristics in chronic tinnitus patients (Y. Chen, Xia, et al., 2015). In addition, enhanced functional connectivity between auditory cortex and the homologue frontal gyrus was observed, indicative of enhanced connectivity of auditory networks with attention networks. (Y. Chen, Zhang, et al., 2015). All these results combined are thought to represent the influence tinnitus has on the networks encompassing memory, attention and emotion. Whether this is causal and which direction the causality goes remains unknown.

In this chapter we describe tinnitus-related changes in functional networks in adults with tinnitus compared to adults without tinnitus. We aim to study whether group-differences can be observed in functional connectivity at the group level and whether performance of an engaging task alters these connectivity patterns.

2 Methods

2.1 Participants

Twenty healthy controls and twenty patients suffering from chronic tinnitus were enrolled in this fMRI study after providing written informed consent, in accordance with the medical ethical committee at the University Medical Center Groningen, the Netherlands. Subjects were recruited from the hospital's tinnitus outpatient clinic as well as from advertisements in local media. None of them had a history of neurological or psychiatric disorders. Subjects' characteristics are listed in Table 1 (Langers, de Kleine, & van Dijk, 2012). Except for the presence of tinnitus in the patient group, all subjects were selected to have normal or near-normal hearing up to 8 kHz. Thresholds were determined in a frequency range of 0.25–16 kHz by means of pure-tone audiometry. Subjects also performed the adaptive categorical loudness scaling (ACALOS)

procedure (Brand & Hohmann, 2002). To characterize the participants' self-reported complaints, all subjects filled out the 14-item Hyperacusis Questionnaire, relating to the attentional, social, and emotional aspects of auditory hypersensitivity (Khalfa et al., 2002).

The tinnitus patients also filled out Dutch translations of questionnaires related to their tinnitus, including the Tinnitus Handicap Inventory (THI) measuring tinnitus severity in daily life (Newman, Jacobson, & Spitzer, 1996), the Tinnitus Reaction Questionnaire (TRQ) that assesses the psychological distress associated with tinnitus (Wilson, Henry, Bowen, & Haralambous, 1991), and the Tinnitus Coping Style Questionnaire (TCSQ) that quantifies effective as well as maladaptive coping strategies (Budd & Pugh, 1996). Patients were asked about the laterality of their tinnitus and the character of the tinnitus sound. Using a modified tinnitus spectrum test, patients were asked to indicate the "likeness" of their tinnitus to narrow-band noises of varying center frequencies.

Table 1. Several group characteristics. Numbers indicate mean ± standard deviation (range). All questionnaire scales were expressed to range from 0-100 for ease of interpretation. The significance of group differences was based on Fisher's exact test (for gender and handedness) or Student's t-test (for all other comparisons), and classified as: ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; ◦$p \geq 0.05$. Numbers do not add up to equal the group size because for some patients not all tinnitus characteristics were obtained.

Group		Healthy controls (n=20)	Tinnitus Patients (n=20)	p
Demographics				
	Gender	16 female, 4 male	12 female, 8 male	°
	Handedness	17 right, 3 left	19 right, 1 left	°
	Age (years)	33 ± 13 (21-60)	46 ± 11 (26-60)	**
Audiometry				
	Average threshold (dB HL)	5 ± 5 (-1 to 18)	8 ± 5 (0 to 23)	°
	Loudness range (dB)	98 ± 8 (84 to 113)	84 ± 14 (56 to 105)	***
Self-reported symptoms				
	Hyperacusis (%)	25 ± 14 (8 - 51)	59 ± 16 (29 - 82)	***
	Depression (%)	8 ± 9 (0 - 31)	27 ± 27 (0 - 100)	**
	Dysthymia (%)	18 ± 14 (0 - 44)	38 ± 24 (0-75)	**
Tinnitus effects				
	Tinnitus Handicap (%)	-	43 ± 22 (6-88)	-
	Tinnitus Reaction (%)	-	38 ± 21 (0-88)	-
	Effective Coping (%)	-	54 ± 14 (17-68)	-
	Maladaptive Coping (%)	-	29 ± 15 (3-60)	-
Tinnitus percept				
	Lateralization	-	8 central, 4 right, 2 left	-
	Type	-	16 steady, 3 pulsatile	-
	Bandwidth	-	10 tone, 7 hiss, 2 ring	-
	Frequency	-	16 high, 2 other, 1 low	-

2.2 Imaging Paradigm

The equipment used for this study was a 3.0-T MR scanner (Philips Intera, Best, the Netherlands), supplied with an 8-channel phased-array (SENSE) transmit/receive head coil. The functional imaging sessions performed on subjects in head-first-supine position in the bore of the system included two runs for the resting-state and two runs for the fixed-state designs. Each resting-state run consisted of a dynamic series of 49 identical high-resolution T2*-sensitive

gradient-echo echo-planar imaging (EPI) volume acquisitions (TR 15.0 s; TA 2.0 s; TE 22 ms; FA 90°; matrix 64 x 64 x 48; resolution 3 x 3 x 3 mm^3; interleaved slice order, no slice gap), whereas fixed-state runs consisted of the same acquisition parameters except the TR that for the fixed-state runs was equal to the TA of 2.0s (continuous scanning) and an ascending slice ordering. In total were 361 volumes acquired.

The acquisition volume was oriented parallel to the Sylvian fissure, and centered approximately on the superior temporal sulci. A sparse, clustered-volume sequence was employed in the (unstimulated) resting-state runs to avoid interference from acoustic scanner noise, while continuous acquisitions were employed in the (stimulated) fixed-state runs to provide higher temporal resolution and power. The scanner coolant pump was turned off during to further reduce noise-levels. In order to control subjects' attentional state, they were asked to focus on the fixation cross throughout the imaging time for resting-state runs, or watch four "emotional" three-minute movie fragments (video and audio) for fixed-state runs, while being scanned simultaneously.

To ensure that the ambient noise did not eliminate the patients' tinnitus, they were asked to rate their tinnitus several times (before, between and after imaging runs) on a 0-10 scale, where 0 indicated no perception of tinnitus at all and 10 indicated the strongest tinnitus imaginable. Ratings did not decrease significantly throughout the session (conversely, some patient's indicated that it systematically increased with stress and fatigue) and never reached zero (2 was the lowest rating).

2.3 Stimuli

MR-compatible electrodynamic headphones (MR Confon GmbH, Magdeburg, Germany; (Baumgart et al., 1998)) were used during the fixed-state runs to deliver sound stimuli from a standard PC with soundcard to participants. Subjects were asked to wear foam earplugs underneath the headset to further diminish the acoustic noise created by the scanner. Visual stimuli were presented by means of a built-in MR-compatible LCD display.

2.4 Preprocessing

Nipype (Gorgolewski et al., 2011) was used to preprocess the fMRI data. Preprocessing consisted of realignment (FSL), artifact rejection (RapidArt) and spatiotemporal filtering to reduce physiological noise (Behzadi, Restom, Liau, & Liu, 2007). Further preprocessing steps included band-pass filtering (0.1-1.0 Hz), registration to an individual high-resolution T1-weighted anatomical image and normalization into MNI stereotaxic space using ANTs (Klein et al., 2009). The data were resampled to a 3-mm isotropic resolution and moderately smoothed using an isotropic 5-mm FWHM Gaussian kernel to improve signal to noise ratio.

2.5 Group analysis

Next, Principal Component Analysis (PCA) was performed to obtain 25 components for each dataset and each subject separately, explaining on average 79.8% of the total variance. These components were subsequently fed into a group Probabilistic Independent Component Analysis as implemented in FSL's MELODIC (v. 3.14) using multi-session temporal concatenation. The obtained networks consisted of spatial maps with their respective time-courses. Subject-specific maps and their time-courses were obtained by back-projection using FSL's Dual Regression. This method uses two stages of regression where in a first stage the individual time-courses are obtained that best fit the group maps. In a second stage, the raw data is regressed against these time-courses to obtain subject-specific component-maps corresponding to the overall group-map.

Importantly, the outputs of stage one of the dual regression —the individual time-courses— were variance normalized before the second stage of the dual regression, allowing for assessment of the shape and amplitude of the component maps. Manual inspection of the time-

courses and maps as well as the (spatial) correlation of the maps with previously reported resting-state networks (Smith et al., 2009) were used to distinguish meaningful components from components containing non-neural signals.

For each of the components, a mixed-effects linear model was composed in AFNIs 3dLME (G. Chen, Saad, Britton, Pine, & Cox, 2013) with factors group (tinnitus vs. controls) and task (fixed-state vs. resting state) and their interaction. The subjects' component maps were fed into this analysis and the significance of the factors was assessed for each component. Further, we assessed the power spectral density (PSD) of the corresponding time-courses.

3 Results

A total of 11 networks were identified as neural in origin (Figure 1). Based on a spatial correlation coefficient with previous networks (Smith et al., 2009) we identified clear sensorimotor networks: two visual networks [Vis1, Vis2], one auditory network [Aud1], a somatosensory/motor network [Sens.Motor] and the cerebellum [Cereb]. In addition, typical resting-state networks were found, such as the limbic system [Limbic], the dorsal attention network [DAN1, DAN2], the default mode network [DMN] and a frontoparietal network [Front.Par].

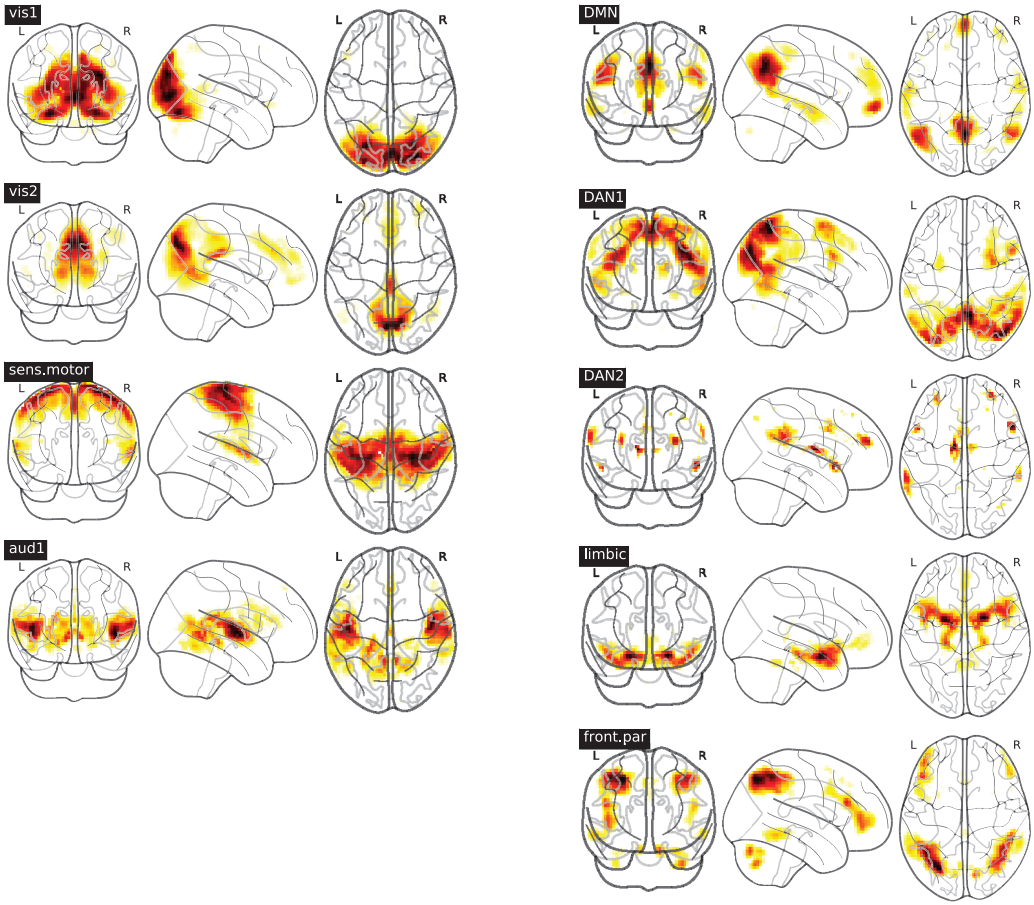


Figure 1. Eleven components obtained from the group analysis using multi-session temporal concatenation of all individual data. Each of the components is thresholded at $z > 3$. The maps in the left column are classical task-related sensory networks while those in the right column show typical resting-state networks.

From these group-level networks, we obtained subject-specific component-maps by applying a dual-regression technique (Filippini et al., 2009). This method uses two stages of regression,

similar to back-projection of individual data to group maps used in other studies. By applying dual-regression we allow for variation between subjects, while maintaining the correspondence of components across subjects.

The subject-specific maps were subsequently fed into a linear mixed model (AFNI's 3dLME) with factors group and task. Importantly, the outputs of stage one of the dual regression —the individual time-courses— were variance normalized before the second stage of the dual regression. In this manner the linear-mixed model now tests for both differences between groups in the shape of the network as well as in the amplitude of the network. The results from these analyses indicated that several networks behaved differently in controls than in subjects with tinnitus. Most notably it identified the default mode network (DMN), roughly encompassed by two distinct but similar networks in the data

Figure 2 shows the two networks that partly overlap. Panel A indicates the DMN that also correlates well with previous demarcations of the DMN (spatial correlation = 0.8) (Smith et al., 2009). In addition, panel E shows a component consisting of a visual area and the posterior and anterior cingulate gyrus. In both the DMN and the latter component, there are adjacent voxels that show a clear and significant group difference ($p < 0.001$ for each of the maps, corresponding to a F-value of 12.71). When looking at the map strength, a proxy for functional connectivity, patients consistently show decreased functional connectivity (panels B and F). Moreover, whereas controls show a clear task-effect, as would be expected for the DMN, this task-effect is much smaller in patients (panel B). For the network with the visual cortex and cingulate gyrus (panel E), this task-effect is less clear (panel F) and shows in controls a decrease during resting state compared to watching a movie, presumably due to the lack of visual stimulation throughout the acquisition of resting-state data. The pattern of reduced connectivity is also clearly visible as a decreased power in the time-courses of the same voxels. This is shown for the both the fixed-state data (panels C and G) and the resting-state data (panels D and F). In all cases, the power is significantly lower in the patient groups.

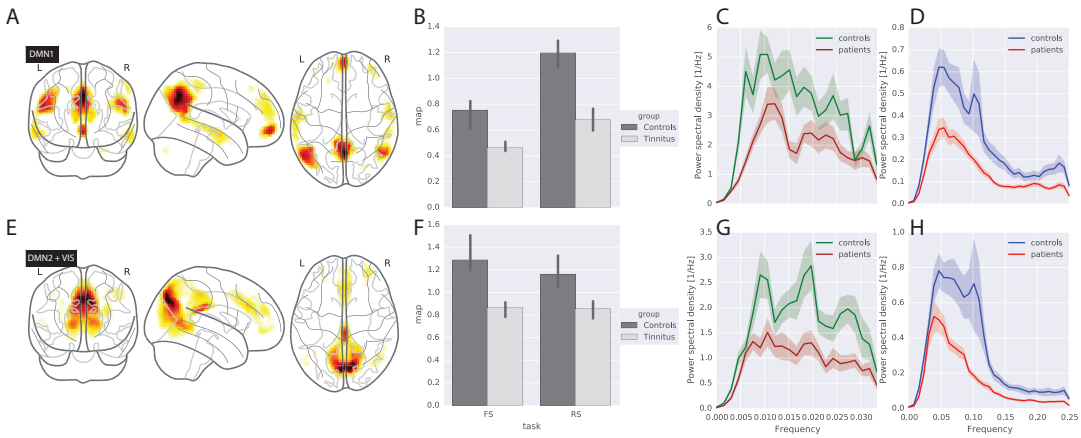


Figure 2. Group differences within two neural networks. The top row shows the default mode network (DMN) whereas the bottom row shows the combination of a visual area and the anterior and posterior cingulate gyrus. Panels B and F show the differences in map strength between the groups and tasks. Panels C and G show the power spectra of the two groups for the fixed-state data whereas panels D and H show the same for the resting state data.

Figure 3 shows the areas where the subject groups differ significantly are almost if not completely adjacent, indicating the potential relevance of the two spatial networks in tinnitus.

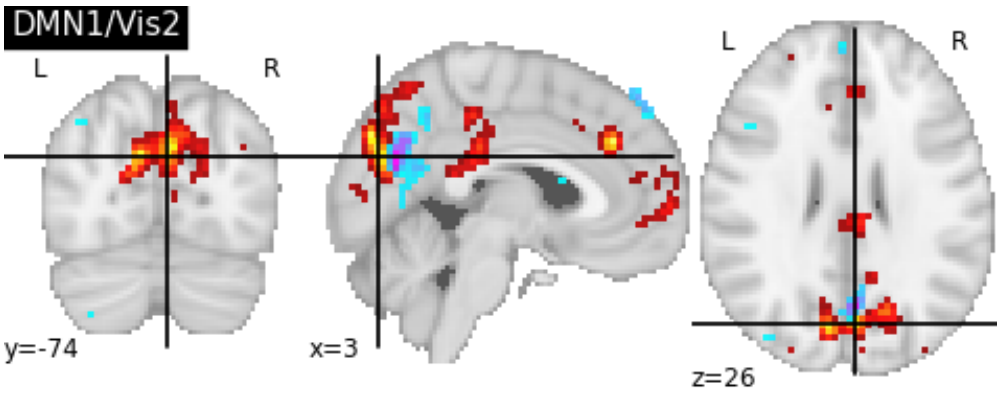


Figure 3. Group differences within two neural networks. The red-yellow colours indicate the significant differences in the network representing a visual area and the cingulate gyrus (see Figure 2E), and the blue-purple colours indicate the differences in the default-mode network (Figure 2A). These areas are remarkably adjacent, corresponding to the left and right (pre)cuneus. This area shows significantly reduced functional connectivity in patients compared to controls in both networks

4 Discussion

This study shows that by using resting-state fMRI in combination with independent component analysis we are able to detect differences in neural networks between substantially-sized subject groups. Resting-state fMRI is a valuable addition to traditional task-related fMRI, where a difference between conditions is obtained. This might not necessarily be the best paradigm to study tinnitus. The consequence of a task-related design is that, by using a subtraction approach, the tinnitus-related activity, present in both conditions, will not be present in the subtraction image unless the task affects the perception of tinnitus (Lanting et al., 2009).

Our data analysis is different from other studies that have used resting-state fMRI to study tinnitus. Often the decomposition of the data into independent component is done by performing two (or more) separate ICA analyses on two (or more) groups (see e.g. (Schmidt, Akrofi, Carpenter-Thompson, & Husain, 2013)). This leads to a difficult interpretation since it is then unclear whether group differences represent (true) differences between subject groups, or whether the data-decomposition was slightly different (biases) for the two groups.

In contrast, we adopted a strategy where the data is pooled over all participants and data-modalities (resting- and fixed-state data). In this way, the data-decomposition is not determined in a potentially different way for both groups, leaving out the necessity of post-ICA matching of corresponding component maps. This allows for a better group-estimate of the components. In addition, by using a dual regression technique we can obtain subject-specific maps and perform group-wise analyses on these maps (Filippini et al., 2009).

This study shows that by using resting-state fMRI in combination with independent component analysis we are able to detect differences in neural networks between substantially-sized subject groups. Resting-state fMRI is a valuable addition to traditional task-related fMRI, where a difference between conditions is obtained. This might not necessarily be the best paradigm to study tinnitus. The consequence of a task-related design is that, by using a subtraction approach, the tinnitus-related activity, present in both conditions, will not be present in the subtraction image unless the task affects the perception of tinnitus (Lanting et al., 2009).

Our data analysis is different from other studies that have used resting-state fMRI to study tinnitus. Often the decomposition of the data into independent component is done by performing two (or more) separate ICA analyses on two (or more) groups (see e.g. Schmidt et al., 2013). This lead to a difficult interpretation since it is then unclear whether group

differences represent (true) differences between subject groups, or whether the data-decomposition was slightly different (biases) for the two groups.

In contrast, we adopted a strategy where the data is pooled over all participants and data-modalities (resting- and fixed-state data). In this way, the data-decomposition is not determined in a potentially different way for both groups, leaving out the necessity of post-ICA matching of corresponding component maps. This allows for a better group-estimate of the components. In addition, by using a dual regression technique we can obtain subject-specific maps and perform group-wise analyses on these maps (Filippini et al., 2009). Using this methodological approach we found a significant group effects for the DMN. This network has been previously implicated in other diseases such as Alzheimer's disease. Lower functional connectivity scores in the DMN were found in an Alzheimer's disease group when compared to normal groups (Binnewijzend et al., 2012).

In contrast, results of resting-state data-analysis in schizophrenia have been variable (M. Greicius, 2008). Schizophrenia is regarded as a large cluster of profound neuropsychiatric symptoms. A subtype of patients (also) experience phantom perception of sounds. One interpretation is that, since the DMN is different between groups in previous work and ours, the DMN somehow plays a role in 'hearing' internally generated sound (whether it is meaningful like in patients with schizophrenia or meaningless like tinnitus).

Our interpretation is that the activity of the DMN in patients with tinnitus, whether as the source or the consequence of tinnitus, is reduced in comparison to the activity in controls. The DMN is relatively more active (i.e. larger spontaneous fluctuations) when the mind is at rest and allowed to wander or daydream. This network is believed to be key in how a person introspectively understands themselves and others, and forms beliefs, intentions, and desires through autobiographical memory. Recently it was shown that a dysfunctional DMN is related to autism (Washington et al., 2014). Perhaps the DMN plays a similar role in tinnitus, where through the on-going tinnitus patients can't get into a state of mind-wandering and therefore show a reduced activity in the DMN compared to controls.

Our study is novel in the sense that it specifically assesses the influence of performing a task on the scale of brain networks as a whole. Specifically in patients with tinnitus, the DMN does not clearly show a task-effect, which in contrast is clearly present in controls. This observation enhances the hypothesis that subjects with tinnitus cannot get the 'rest' that is necessary for a healthy life. In conclusion, we attribute the pattern of reduced connectivity in the DMN to the unremitting engaging effect of the tinnitus percept.

Acknowledgments

DRML was funded by VENI research grant 016.096.011 from the Netherlands organization for scientific research (NWO) and the Netherlands organization for health research and development (ZonMw). Further financial support was provided by the Heinsius Houbolt Foundation.

References

Adjamian, P., Hall, D. A., Palmer, A. R., Allan, T. W., & Langers, D. R. M. (2014). Neuroanatomical abnormalities in chronic tinnitus in the human brain. *Neuroscience and Biobehavioral Reviews*. doi:10.1016/j.neubiorev.2014.05.013

Baumgart, F., Kaulisch, T., Tempelmann, C., Gaschler-Markefski, B., Tegeler, C., Schindler, F., … Scheich, H. (1998). Electrodynamic headphones and woofers for application in magnetic resonance imaging scanners. *Medical Physics*, *25*(10), 2068–70.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. doi:10.1016/j.neuroimage.2007.04.042

Binnewijzend, M. A. A., Schoonheim, M. M., Sanz-Arigita, E., Wink, A. M., van der Flier, W. M., Tolboom, N., … Barkhof, F. (2012). Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment. *Neurobiology of Aging*, *33*(9), 2018–2028. doi:10.1016/j.neurobiolaging.2011.07.003

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, *112*(4), 1597–1604. doi:10.1121/1.1502902

Budd, R. J., & Pugh, R. (1996). Tinnitus coping style and its relationship to tinnitus severity and emotional distress. *Journal of Psychosomatic Research*, *41*(4), 327–335. doi:10.1016/S0022-3999(96)00171-7

Burton, H., Wineland, A., Bhattacharya, M., Nicklaus, J., Garcia, K. S., & Piccirillo, J. F. (2012). Altered networks in bothersome tinnitus: a functional connectivity study. *BMC Neuroscience*, *13*(1), 3. doi:10.1186/1471-2202-13-3

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage*, *73*, 176–90. doi:10.1016/j.neuroimage.2013.01.047

Chen, Y., Xia, W., Feng, Y., Li, X., Zhang, J., Feng, X., … Teng, G. (2015). Altered Interhemispheric Functional Coordination in Chronic Tinnitus Patients. *BioMed Research International*, *2015*, 1–8. doi:10.1155/2015/345647

Chen, Y., Zhang, J., Li, X., Xia, W., Feng, X., Qian, C., … Teng, G. (2015). Altered Intra- and Interregional Synchronization in Resting-State Cerebral Networks Associated with Chronic Tinnitus. *Neural Plasticity*, *2015*, 1–11. doi:10.1155/2015/475382

Davies, J., Gander, P. E., Andrews, M., & Hall, D. a. (2014). Auditory network connectivity in tinnitus patients: a resting-state fMRI study. *International Journal of Audiology*, *53*(3), 192–8. doi:10.3109/14992027.2013.846482

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., … Mackay, C. E. (2009). Distinct patterns of brain activity in young carriers of the APOE-epsilon4 allele. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(17), 7209–14. doi:10.1073/pnas.0811879106

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*(August), 13. doi:10.3389/fninf.2011.00013

Greicius, M. (2008). Resting-state functional connectivity in neuropsychiatric disorders. *Current Opinion in Neurology*, *21*(4), 424–430. doi:10.1097/WCO.0b013e328306f2c5

Greicius, M. D., Srivastava, G., Reiss, A. L., & Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(13), 4637–42. doi:10.1073/pnas.0308627101

Khalfa, S., Dubal, S., Veuillet, E., Perez-Diaz, F., Jouvent, R., & Collet, L. (2002). Psychometric normalization of a hyperacusis questionnaire. *ORL*, *64*, 436–442. doi:10.1159/000067570

Kim, J., Kim, Y., Lee, S., Seo, J.-H., Song, H.-J., Cho, J. H., & Chang, Y. (2012). Alteration of functional connectivity in tinnitus brain revealed by resting-state fMRI?: A pilot study. *International Journal of Audiology*. doi:10.3109/14992027.2011.652677

Klein, A., Andersson, J., Ardekani, B. a, Ashburner, J., Avants, B., Chiang, M.-C., … Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, *46*(3), 786–802. doi:10.1016/j.neuroimage.2008.12.037

Langers, D. R. M., de Kleine, E., & van Dijk, P. (2012). Tinnitus does not require macroscopic tonotopic map reorganization. *Frontiers in Systems Neuroscience*, *6*, 2. doi:10.3389/fnsys.2012.00002

Lanting, C. P., de Kleine, E., & van Dijk, P. (2009). Neural activity underlying tinnitus generation: results from PET and fMRI. *Hearing Research*, *255*(1-2), 1–13. doi:10.1016/j.heares.2009.06.009

Maudoux, a, Lefebvre, P., Cabay, J.-E., Demertzi, a, Vanhaudenhuyse, a, Laureys, S., & Soddu, a. (2012). Connectivity graph analysis of the auditory resting state network in tinnitus. *Brain Research*, *1485*, 10–21. doi:10.1016/j.brainres.2012.05.006

Maudoux, A., Lefebvre, P., Cabay, J.-E., Demertzi, A., Vanhaudenhuyse, A., Laureys, S., & Soddu, A. (2012). Auditory resting-state network connectivity in tinnitus: a functional MRI study. *PloS One*, *7*(5), e36222. doi:10.1371/journal.pone.0036222

Newman, C. W., Jacobson, G. P., & Spitzer, J. B. (1996). Development of the Tinnitus Handicap Inventory. *Archives of Otolaryngology--Head & Neck Surgery*, *122*, 143–148. doi:10.1001/archotol.1996.01890140029007

Schmidt, S. a, Akrofi, K., Carpenter-Thompson, J. R., & Husain, F. T. (2013). Default mode, dorsal attention and auditory resting state networks exhibit differential functional connectivity in tinnitus and hearing loss. *PloS One*, *8*(10), e76488. doi:10.1371/journal.pone.0076488

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., … Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 13040–5. doi:10.1073/pnas.0905267106

Washington, S. D., Gordon, E. M., Brar, J., Warburton, S., Sawyer, A. T., Wolfe, A., … Vanmeter, J. W. (2014). Dysmaturation of the default mode network in autism. *Human Brain Mapping*, *35*(4), 1284–1296. doi:10.1002/hbm.22252

Wilson, P. H., Henry, J., Bowen, M., & Haralambous, G. (1991). Tinnitus reaction questionnaire: psychometric properties of a measure of distress associated with tinnitus. *Journal of Speech and Hearing Research*, *34*(1), 197–201.

Zhang, D., Snyder, A. Z., Fox, M. D., Sansbury, M. W., Shimony, J. S., & Raichle, M. E. (2008). Intrinsic functional relations between human cerebral cortex and thalamus. *Journal of Neurophysiology*, *100*(4), 1740–1748. doi:10.1152/jn.90463.2008

Chapter 20

The Role of Conduction Delay in Creating Sensitivity to Interaural Time Differences

Catherine E. Carr, Go Ashida
Department of Biology, University of Maryland, College Park, MD, USA

Hermann Wagner
Institute for Biology II, RWTH Aachen University, Aachen, Germany

Thomas McColgan, Richard Kempter
Humboldt University, Berlin, Germany

Contact author: CEC, cecarr@umd.edu

Abstract

Axons from the nucleus magnocellularis (NM) and their targets in nucleus laminaris (NL) form the circuit responsible for encoding interaural time difference (ITD). In barn owls, NL receives bilateral inputs from NM, such that axons from the ipsilateral NM enter NL dorsally, while contralateral axons enter from the ventral side. These afferents act as delay lines to create maps of ITD in NL. Because NL is tonotopically organized, these maps are technically maps of interaural phase difference; time difference is computed in the inferior colliculus after combining across frequency channels (Wagner, Takahashi, & Konishi, 1987). Since delay Iine inputs are characterized by a precise latency to auditory stimulation, but the postsynaptic coincidence detectors respond to ongoing phase difference, we asked whether the latencies of a local group of axons were identical, or varied by multiples of the inverse of the frequency they respond to, which is equivalent to multiples of 2π phase. Intracellular recordings from NM axons were used to measure delay-line latencies in NL. Systematic shifts in conduction delay within NL accounted for the maps of ITD, but recorded latencies of individual inputs at nearby locations varied by 2π or 4π. Therefore microsecond precision is achieved through sensitivity to phase delays, rather than absolute latencies. We propose that the auditory system "coarsely" matches ipsilateral and contralateral latencies using physical delay lines, so that inputs arrive at NL at about the same time, and then "finely" matches latency modulo 2π to achieve microsecond ITD precision.

Keywords: coding, interaural time difference, plasticity, models, rate

1 Introduction

A key feature of ITD sensitivity is how the observed microsecond delays may be created by neural elements that are both noisy and slow. There are two possibilities, illustrated by the example of a coincidence detector neuron that responds to a pure sine tone, presented to both ears at an ITD of 0 μs, or sound in front of the animal. One strategy requires equal conduction latencies from each ear to the coincidence detector, such that the inputs would arrive with equal latency and equal phase. The second strategy relies on relative delays and would simply require

that inputs arrive at the same phase. Due to the periodicity of the sine wave, both scenarios would produce the same output. More realistic inputs would not be perfectly periodic, but the coding strategy remains viable because of the tonotopic organization of the auditory system, which generates ringing properties of the inputs that are characteristic of the particular frequency (Carney, McDuffy, & Shekhter, 1999; Recio et al., 1998; Wagner et al., 2005). Previously we had shown that the inputs from the cochlear nuclei act as delay line inputs to form maps of ITD in the barn owl (Carr & Konishi, 1990), but could not determine whether the delay lines had identical conduction latencies or were matched for phase only.

To differentiate between the two hypotheses, we combined measurements of the extracellular potential or neurophonic with intracellular measures of click latency in NL. For mapping the ITD we used the neurophonic, because it reflects local ITD sensitivity (Kuokkanen et al., 2013; Kuokkanen et al., 2010; Mc Laughlin, Verschooten, & Joris, 2010; Sullivan & Konishi, 1986), and to measure delay we used clicks, because their temporal occurrence is precise to within about 20μs (Wagner et al., 2005). We found the microsecond precision needed to construct maps of ITD is achieved through precisely regulated phase delays, rather than by regulation of absolute latency, consistent with the observation that nucleus laminaris neurons respond over a wide range of integer multiples of their preferred interaural phase difference (Christianson & Pena, 2006). Thus latencies from ear to the point of coincidence detection need only be adjusted within a stimulus period, creating a flexible yet precise system for detection of sound source sound location. This high precision approach should also benefit other binaural properties of the auditory system, such as resistance to background noise and reverberation.

2 Methods

The experiments were conducted at the University Maryland. Four barn owls (*Tyto alba*) were used to collect these data, and procedures conformed to NIH guidelines for Animal Research and were approved by the Animal Care and Use Committee of the Universities of Maryland. Anaesthesia was induced by intramuscular injections of 10-20mg/kg ketamine hydrochloride and 3-4mg/kg xylazine, with supplementary doses to maintain a suitable plane of anaesthesia. Recordings were made in a sound-attenuating chamber (IAC, New York). Microelectrodes filled with 2 M K-acetate were used for intracellular recordings, and amplified by an Axoclamp 2B, and then by a spike conditioner (PC1, Tucker-Davis Technologies (TDT), Gainesville, FL). Acoustic stimuli were digitally generated by custom-made software ("Xdphys" written in Dr. M. Konishi's lab at Caltech) driving a signal-processing board (DSP2 (TDT)). Acoustic signals were fed to earphones, inserted into the owl's left and right ear canals, respectively. At a given recording site, we measured frequency tuning, then tuning to ITD, and responses to monaural clicks.

3 Results

Axons from the cochlear nucleus magnocellularis (NM) form presynaptic maps of interaural time differences (ITD) in the nucleus laminaris (NL). These inputs generate a field potential that varies systematically with recording position and can be used to map ITDs. In the barn owl, the representation of best ITD shifts systematically in NL, forming multiple, largely overlapping maps of ITD. We had previously found that conduction delays could account for the shift in maps of ITD (Carr et al., 2013).

3.1 Measuring latency with click delays

Our mapping studies showed that the representation of frontal space (0 μs ITD) in NL shifted systematically to more ventral locations with the progression from medial to lateral along each tonotopic slab (red line, Fig. 1A, see Carr et al., 2013). To determine if changes in conduction velocity formed the basis of the systematic shift, we used click delay to measure latencies at

different mediolateral positions. Measurements of click latency provided reliable measures of conduction delay (Köppl & Carr, 2008; Wagner et al., 2005), because click stimuli were temporally precise. Nevertheless, click stimuli evoked an oscillatory response (Ruggero et al., 1986; Wagner et al., 2005; Wagner et al., 2009), in which typically several peaks and troughs could be distinguished (Fig. 1B).

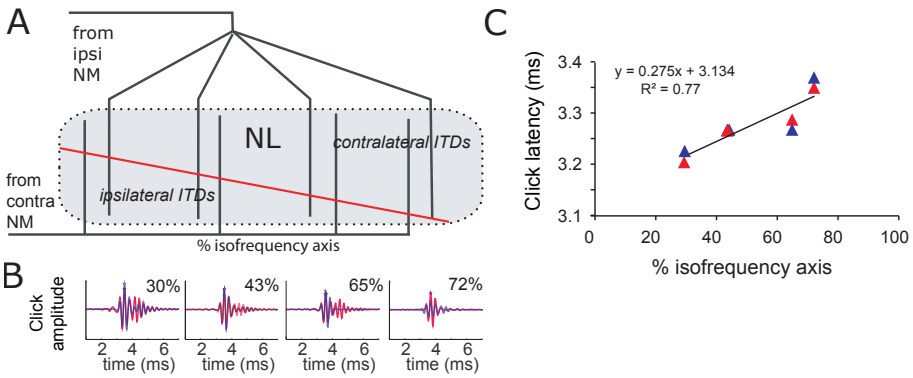


Figure 1. A. Schematic outline of NL showing delay line inputs from each side, with the "0µs-ITD line" in red. B. Click responses were recorded best ITDs around 0µs, at four different locations along a 3.6 kHz iso-frequency slab in NL. Superimposed ipsilateral (blue) and contralateral (red) averages of 128 click responses from recording sites at 30, 43, 65 and 72% along the isofrequency axis. C. Click delays from B. showed a systematic increase in latency with mediolateral position in NL for both ipsi- and contralateral responses.

Neurophonic responses to ipsi- and contralateral clicks at best ITDs at or near 0 µs generated similar but not identical ipsi- and contralateral click responses, which largely overlapped (Fig. 1B). Responses to both ipsilateral and contralateral stimulation showed a steady increase in latency along the iso-frequency axis in NL (Fig. 1C), i.e. systematic changes in conduction delay underlay the formation of the maps of ITD. To determine whether the delay lines had identical conduction latencies or were matched for phase only, we compared neurophonic multiunit click recordings with intracellular recordings from individual NM axons in NL.

For well-isolated extracellular single-unit recordings, PSTH histograms of spike responses to monaural click stimuli can be used to measure delay in, for example, the auditory nerve and the nucleus magnocellularis (Köppl, 1997; Köppl & Carr, 2008). This approach could not be used for the extracellular neurophonic, which represented the sum of many (>250) afferent sources, and which was characterized by robust oscillations at a frequency close to BF that overwhelmed single-unit responses (Köppl & Carr, 2008; Kuokkanen et al., 2010; see Fig. 2, insert). We therefore measured intracellular responses from NM axons in NL to click stimuli in 3 owls (n=43 units, 2.6 - 5.2 kHz best-frequency regions of NL) to determine if individual response delays reflected the delays of the peaks and troughs of the neurophonic click response.

Intracellular recordings from NM axons in NL supported the hypothesis that oscillatory neurophonic click responses were composed of many NM click responses. Averaging intracellular click recordings from NM afferents over hundreds of repetitions revealed two to five PSTH peaks (Fig. 2). PSTH peaks were separated in time by intervals equal to the inverse of the neuron's best frequency (2π phase interval); a phase difference of 4π between the first and second peak was observed in one case (Kiang, 1967). Individual spike timing varied, consistent with phase locking (Anderson et al., 1970; Joris et al., 1994; Köppl, 1997; Sachs, Young, & Lewis, 1974).

PSTH peaks derived from intracellular recordings of NM axons resembled half wave rectified neurophonic extrema (Fig. 2, 3A). For each intracellular recording, we obtained responses to 300 - 1,000 repetitions, and quantified the temporal distance between the PSTH peaks. All showed similar 2π separation of peaks. Adjacent recordings responded at very similar phases,

suggesting adjacent axons fire in synchrony or at multiples of 2π (Fig. 3B, note clustered adjacent symbols).

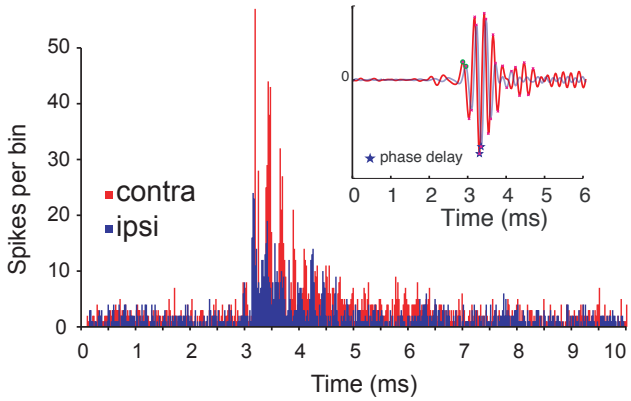


Figure 2. Intracellular and extracellular click responses in NL. Both neurophonic responses to stimulation of ipsi- and contralateral ears (insert), and intracellular recordings of two adjacent NM afferents in NL, one ipsi- and the other contralateral, revealed PSTHs with multiple phase locked peaks. Both intracellular responses had peaks separated by about 192 μs apart, or 1/BF where BF=5.2 kHz, i.e by 2π.

3.2 Latencies of adjacent recordings can vary by multiples of 2π

In the penetration through the dorsoventral dimension of NL shown in Fig. 3, recordings between 205- 270 μm depths yielded three units with similar phase, but a latency difference of 4π (Fig. 3C, black arrow). The first two recordings had latencies of 3.33 and 3.32 ms, the third 2.91 ms. Thus the recording with the shorter latency occurred almost 4π earlier than the first two. In other words, the phase delay of these peaks exhibited a narrow scatter, while the signal front delay showed a large scatter, as had also been observed for the neurophonic (Wagner et al., 2005). A similar measurement at 865 μm had a latency of 3.03, while the adjacent recordings at 725 and 925 μm had latencies of 3.21 and 3.20 ms (Fig. 3C, grey arrow). Most measurements had similar phase *and* latency (Fig. 3D, bar marked 0), but some occurred 2π earlier, and others 2π or 4π later than the rest of the population (Fig. 3D).

In general, changes in delay with depth were consistent with the NM axons acting as delay lines (Carr & Konishi, 1990), and with modelled conduction velocities in (McColgan et al., 2014). Ipsilateral recordings showed an increase in latency with recording depth in NL (Fig. 3B, C), while contralateral recordings showed a decrease in latency with depth, consistent with the delay line geometry. The regression lines in Fig. 3C predicted latencies of 2.5-8.3 m/s, consistent with predicted changes in delay (McColgan et al., 2014) and measurements from previous studies (Carr & Konishi, 1990). Nevertheless, we point out that while these changes in delay would be expected from multiple recordings from the same NM axon, they were not required for ITD sensitivity. Sensitivity to ITDs should be instead conferred by the differences in phase present.

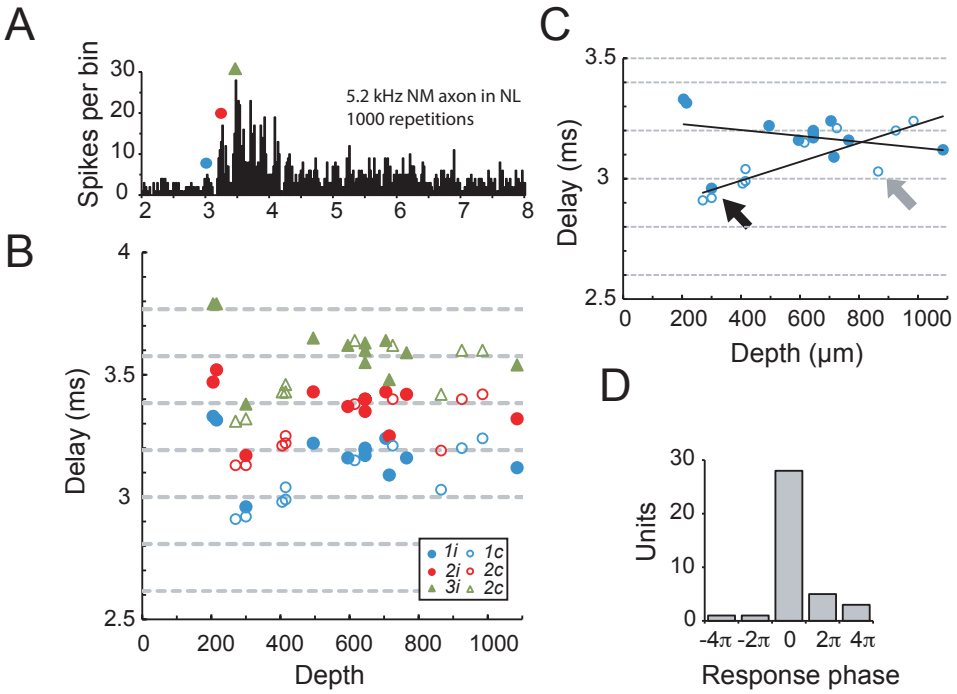


Figure 3. A. Delay of the first three PSTH peaks recorded in a single NM axon within NL. Coloured symbols mark peaks in the PSTH. Blue circle = first peak, or signal front delay, red circle = second peak and green triangle = third peak. B. Click latencies in 21 NM afferents recorded in a single penetration through the 5 kHz region of left NL. Both ipsilateral (open symbols) and contralateral (closed symbols) responses were encountered, symbols as in A. The latencies of the first 3 PSTH peaks for these afferents varied by multiples of 1/BF, shown as horizontal dashed lines 190 μs apart; BFs varied from 4.4-5.2 kHz. The first three PSTH peaks for each recording were shown with respect to recording depth as a vertical array of blue squares (first peak), red circles (second peak, +2π) and green triangles (third peak, +4π). C. Latencies of the ipsilateral NM axons increased with depth (open symbols, y=0.4 s/m x + 2.84ms, r^2 =0.7), while recordings from contralateral NM axons in NL decreased in latency with depth (filled symbols, y = -0.12 s/m x + 3.25, r^2 = 0.10. D. Data from 3 penetrations through NL (n=38 units, 2 owls) showed systematic shifts in latency with recording depth, consistent with models of delay (0μs bin). Other units were 2π or 4π earlier or later than neighbouring units (C, arrows).

Based on NL architecture and our results, we have generated an explanation of how conduction velocity of the incoming delay lines axons from NM affects maps of ITD in NL. Both extracellular neurophonic recordings and intracellular NM axonal recordings in NL revealed maps of best ITD, created by precisely regulated conduction delays from intrinsically noisy elements. The microsecond precision needed to construct these maps of ITD appears to be achieved through precisely regulated phase delays, rather than by regulation of absolute latency.

4 Discussion

Delays from each side must be matched for coincidence detection to occur. How? Coincidence requires microsecond precision, which may be achieved by a two-step process, where latencies from each ear are matched during development, and then precisely matched modulo 2π. This was first proposed by Gerstner at al. (1996), who postulated that the necessary degree of coherence in the signal arrival times could be attained during development by an unsupervised Hebbian learning rule that selects connections with matching delays from a broad distribution of axons with random delays. Our data support this matching hypothesis, with coarse delays laid

down through normal development, and precise, i.e. ± 20 μs modulo 2π, delays regulated by some as yet unknown activity dependent processes (Seidl, Rubel, & Harris, 2010).

Physiological data are consistent with coarsely matched latencies. For each frequency, latencies in barn owl NM and in the midline cross fibre tract are similar, consistent with a common path from the ear into the brain (see Fig. 7 in Köppl, 1997). Our previous measures of latency did not have the precision of the current intracellular recordings, with ipsilateral and contralateral delays at the dorsal and ventral borders of NL of 2.82 ± 0.24 and 2.87 ± 0.21 ms respectively (Carr & Konishi, 1990). Note the standard deviations of about 200 μs in these early recordings of latency. Our current recordings support the hypothesis that within NL, latencies vary by multiples of 2π, and may be precisely regulated at a fine time scale, in order to create a cycle by cycle representation of the stimulus at the point(s) of coincidence detection (Ashida, Funabiki, & Carr, 2013; Funabiki, Ashida, & Konishi, 2011). Modulation by multiples of 2π is also consistent with cross-correlation and spike timing dependent plasticity models (Fischer, Christianson, & Pena, 2008; Gerstner et al., 1996; Kempter et al., 1998; Pena & Konishi, 2000). Variability in response latency also characterizes mammalian auditory nerve and cochlear nucleus recordings (Carney & Yin, 1988; Sanes & Constantine-Paton, 1985; Young, Robert, & Shofner, 1988).

Acknowledgments

This research was sponsored by NIH DC00436 to CEC, by NIH P30 DC04664 to the University of Maryland Center for the Comparative and Evolutionary Biology of Hearing, by the German Research Foundation (DFG, Wa-606/12, Ke-788/1-3, 4) and the Bundesministerium für Bildung und Forschung (BMBF, Bernstein Collaboration Temporal Precision, 01GQ07101 to HW and 01GQ07102 to RK), and by the cluster of Excellence, "Hearing4all" at the University of Oldenburg (GA).

References

Anderson DJ, Rose JE, Hind, J. E., & Brugge. (1970). Temporal Position of Discharges in Single Auditory Nerve Fibers within the Cycle of a Sine-Wave Stimulus: Frequency and Intensity Effects. *The Journal of the Acoustical Society of America*, *49(2)*, 1131–1139.

Ashida, G., Funabiki, K., & Carr, C. E. (2013). Biophysical basis of the sound analog membrane potential that underlies coincidence detection in the barn owl. *Frontiers in Computational Neuroscience*, *7*, 102. doi:10.3389/fncom.2013.00102

Carney, L. H., & Yin, T. C. T. (1988). temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *Journal of Neurophysiology*, *60*, 1653–1677.

Carney, L., McDuffy, M., & Shekhter, I. (1999). Frequency glides in the impulse responses of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, *105*, 2384.

Carr, C. E., & Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *The Journal of Neuroscience*, *10*(10), 3227–3246.

Carr, C. E., Shah, S., Ashida, G., McColgan, T., Wagner, H., Kuokkanen, P. T., et al. (2013). Maps of ITD in the nucleus laminaris of the barn owl. *Advances in Experimental Medicine and Biology*, *787*, 215–222. doi:10.1007/978-1-4614-1590-9_24

Christianson, G. B., & Pena, J. L. (2006). Noise reduction of coincidence detector output by the inferior colliculus of the barn owl. *The Journal of Neuroscience*, *26*(22), 5948–5954.

Fischer, B. J., Christianson, G. B., & Pena, J. L. (2008). Cross-correlation in the auditory coincidence detectors of owls. *The Journal of Neuroscience*, *28*(32), 8107–8115.

Funabiki, K., Ashida, G., & Konishi, M. (2011). Computation of Interaural Time Difference in the Owl's Coincidence Detector Neurons. *The Journal of Neuroscience*, *31*(43), 15245–15256.

Funabiki, K., Koyano, K., & Ohmori, H. (1998). The role of GABAergic inputs for coincidence detection in the neurones of nucleus laminaris of the chick. *The Journal of Physiology*, *508*, 851–869.

Gerstner, W., Kempter, R., van Hemmen, J. L., & Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, *383*(6595), 76–81. doi:10.1038/383076a0

Joris, P. X., Carney, L. H., Smith, P. H., & Yin, T. C. T. (1994). Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency. *Journal of Neurophysiology*, *71*(3), 1022–1036.

Kempter, R., Gerstner, W., van Hemmen, J. L., & Wagner, H. (1998). Extracting oscillations. Neuronal coincidence detection with noisy periodic spike input. *Neural Computation*, *10*(8), 1987–2017.

Kiang, N. (1967). Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. *Journal of Anatomy*, *101*(Pt 1), 176.

Köppl, C. (1997). Frequency tuning and spontaneous activity in the auditory nerve and cochlear nucleus magnocellularis of the barn owl Tyto alba. *Journal of Neurophysiology*, *77*, 364–377.

Köppl, C., & Carr, C. E. (2008). Maps of interaural time difference in the chicken's brainstem nucleus laminaris. *Biological Cybernetics*, *98*(6), 541–559. doi:10.1007/s00422-008-0220-6

Kuokkanen, P. T., Ashida, G., Carr, C. E., Wagner, H., & Kempter, R. (2013). Linear summation in the barn owl's brainstem underlies responses to interaural time differences. *Journal of Neurophysiology*, *110*(1), 117–130. doi:10.1152/jn.00410.2012

Kuokkanen, P. T., Wagner, H., Ashida, G., Carr, C. E., & Kempter, R. (2010). On the origin of the extracellular field potential in the nucleus laminaris of the barn owl (Tyto alba). *Journal of Neurophysiology*, *104*(4), 2274–2290. doi:10.1152/jn.00395.2010

Mc Laughlin, M., Verschooten, E., & Joris, P. X. (2010). Oscillatory dipoles as a source of phase shifts in field potentials in the mammalian auditory brainstem. *The Journal of Neuroscience*, *30*(40), 13472–13487. doi:10.1523/JNEUROSCI.0294-10.2010

McColgan, T., Shah, S., Köppl, C., Carr, C. E., & Wagner, H. (2014). A functional circuit model of interaural time difference processing. *Journal of Neurophysiology*, *112*(11), 2850–2864. doi:10.1152/jn.00484.2014

Pena, J. L., & Konishi, M. (2000). Cellular mechanisms for resolving phase ambiguity in the owl's inferior colliculus. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11787–11792. doi:10.1073/pnas.97.22.11787

Recio, A., Rich, N., Narayan, S., & Ruggero, M. (1998). Basilar-membrane responses to clicks at the base of the chinchilla cochlea. *The Journal of the Acoustical Society of America*.

Ruggero, M., Robles L, Rich, N., & Costalupes, J. (1986). Basilar membrane motion and spike Initiation in the cochlear nerve. In B. C. J. Moore & R. D. Patterson, *Auditory Frequency Selectivity* (119 ed., Vol. 189). London: Auditory Frequency Selectivity.

Sachs, M., Young, E., & Lewis, R. (1974). Discharge patterns of single fibers in the pigeon auditory nerve. *Brain Research*, *70*(3), 431–447.

Sanes, D. H., & Constantine-Paton, M. (1985). The sharpening of frequency tuning curves requires patterned activity during development in the mouse, Mus musculus. *The Journal of Neuroscience*, *5*(5), 1152–1166.

Seidl, A. H., Rubel, E. W., & Harris, D. M. (2010). Mechanisms for adjusting interaural time differences to achieve binaural coincidence detection. *The Journal of Neuroscience*, *30*(1), 70–80. doi:10.1523/JNEUROSCI.3464-09.2010

Sullivan, W. E., & Konishi, M. (1986). Neural map of interaural phase difference in the owl's brainstem. *Proceedings of the National Academy of Sciences of the United States of America*, *83*, 8400–8404.

Wagner, H., Brill, S., Kempter, R., & Carr, C. E. (2005). Microsecond precision of phase delay in the auditory system of the barn owl. *Journal of Neurophysiology*, *94*(2), 1655–1658.

Wagner, H., Brill, S., Kempter, R., & Carr, C. E. (2009). Auditory responses in the barn owl's nucleus laminaris to clicks: impulse response and signal analysis of neurophonic potential. *Journal of Neurophysiology*, *102*(2), 1227–1240. doi:10.1152/jn.00092.2009

Wagner, H., Takahashi, T., & Konishi, M. (1987). Representation of interaural time difference in the central nucleus of the barn owl's inferior colliculus. *The Journal of Neuroscience*, *7*(10), 3105–3116.

Young, E., Robert, J., & Shofner, W. (1988). Regularity and latency of units in ventral cochlear nucleus: implications for unit classification and generation of response properties. *Journal of Neurophysiology*, *60*(1), 1–29.

Chapter 21

Objective Measures of Neural Processing of Interaural Time Differences

David McAlpine, Nicholas Haywood, Jaime Undurraga, Torsten Marquardt
Ear Institute, University College London, London, UK

Contact author: DM, d.mcalpine@ucl.ac.uk

Abstract

We developed an objective measure to assess neural sensitivity to interaural time differences (ITDs) conveyed in the temporal fine structure (TFS) of low-frequency sounds, and to distinguish it from neural sensitivity to ITDs conveyed in the temporal envelope of amplitude-modulated (AM'ed) high-frequency sounds. Bilateral cochlear implant (CI) users can potentially access the latter cues, but are largely unable to process low-frequency ITD information. To this end, a measure of brain activity that distinguishes between neural pathways coding ITDs in the TFS and the stimulus envelope would be useful in developing stimulus strategies to improve spatial listening in CI users. Using electroencephalography (EEG), we recorded brain activity to sounds in which the interaural phase difference (IPD) of the TFS (or the modulated temporal envelope) was repeatedly switched between leading in one ear or the other. When the amplitude of the tones is modulated equally in the two ears at 41 Hz, the interaural phase modulation (IPM) generates an intracranial percept of a sound moving from one side to the other – and evokes an IPM following-response (IPM-FR) in the EEG signal. For low-frequency signals (520-Hz carrier), IPM-FRs were reliably obtained for a wide range of modulation rates and IPDs, and were largest for an IPM rate of 6.8 Hz and when IPD switches (around 0°) were in the range 45-90°**.** Increasing the modulation frequency increased the magnitude of IPM-FRs. IPDs conveyed in envelope of high-frequency tones (3-kHz tones AM'ed at 128 Hz, with a second-order modulation of 41 Hz) also generated a reliable pattern of IPM-FRs, but one in which response maxima occurred for IPDs switched between 0° and 180° IPD. The data are consistent with the interpretation that distinct binaural mechanisms generate the IPM-FR at low and high frequencies, and with the reported physiological responses of medial superior olive (MSO) and lateral superior olive (LSO) neurons in other mammals. Low-frequency binaural neurons in the MSO are considered maximally activated by IPDs in the range 45°-90°, consistent with their reception of excitatory inputs from both ears. High-frequency neurons in the LSO receive excitatory and inhibitory input from the two ears receptively – as such maximum activity occurs when the sounds at the two ears are presented out of phase.

Keywords: interaural phase modulation, objective measures, frequency-following response, electroencephalography

1 Introduction

1.1 Advantages of Binaural Listening

Binaural hearing confers considerable advantages in everyday listening environments. By comparing the timing and intensity of a sound at each ear, listeners can locate its source on the horizontal plane, and exploit these differences to hear out signals in background noise – an important component of 'cocktail party listening'. Sensitivity to interaural time differences (ITDs) in particular has received much attention, due in part to the exquisite temporal performance observed; for sound frequencies lower than about 1.3 kHz, ITDs of just a few tens of microseconds are discriminable, observed both at the neural and behavioural levels. Sensitivity to ITDs decreases with ageing (Herman et al., 1977; Abel et al., 2000; Babkoff et al., 2002), and is typically impaired in hearing loss, albeit variability across individuals and aetiologies (see Moore et al., 1991, for review), with reduced sensitivity most apparent when trying to localise sources in background noise (Lorenzi et al., 1999). ITD processing may also remain impaired following surgery to correct conductive hearing loss (Wilmington et al., 1994), and sensitivity to ITDs is typically poor in users of hearing technologies such as bilateral cochlear implants (CIs). In terms of restoring binaural function in therapeutic interventions, there is considerable benefit to be gained from developing objective measures to assess binaural processing.

1.2 Objective Measures of Binaural Hearing

Direct measures of ITD sensitivity, attributed to multiple generator sites in the thalamus and auditory cortex (Picton et al., 1974; Hari et al., 1980; Näätänen and Picton, 1987; Liégeois-Chauvel et al., 1994), have been demonstrated in auditory-evoked 'P1-N1-P2' responses, with abrupt changes in either the ITD (e.g. McEvoy, 1991), or the interaural correlation of a noise (e.g. Chait et al., 2005), eliciting responses with typical latencies of 50, 100, and 175 ms, respectively. Although these studies demonstrate the capacity to assess binaural processing of interaural temporal information, they provide for only a relatively crude measure of ITD sensitivity, comparing, for example, EEG responses to diotic (identical) sounds at the two ears with responses to sounds that are statistically independent at the two ears. A more refined measure of ITD sensitivity would provide the possibility of assessing neural mechanisms of binaural processing in the human brain, aiding bilateral fitting of CIs in order to maximise binaural benefit.

2 Methods

Eleven NH listeners took part in the low-frequency EEG experiment (6 male; mean age = 24.4 years, range = 18-34 years). Four listeners took part in the high-frequency experiment (3 male; mean age = 30.0 years, range = 27-34 years). All subjects demonstrated hearing levels of 20 dB hearing level or better at octave frequencies between 250-8000 Hz, and reported a range of musical experience. Subjects were recruited from poster advertisements. All experiments were approved by the UCL Ethics Committee. Subjects provided informed consent, and were paid an honorarium for their time. For EEG recordings assessing sensitivity to ITDs in the low-frequency TFS, 520-Hz carrier tones were modulated with 100% sinusoidal AM (SAM), at a rate of 41 Hz. Stimuli were presented continuously for 4 min and 48 s (70 epochs of 4.096 s) at 80 dB SPL. Carrier and modulation frequencies were set so that an integer number of cycles fitted into an epoch window. The carrier was presented with an IPD (in the range ±11.25° to ±135° IPD around 0°), whilst the modulation envelope remained diotic at all times. In order to generate IPM, the overall magnitude of the carrier IPD was held constant throughout the stimulus, but the ear at which the signal was leading in phase was periodically modulated between right and left. Each IPM was applied instantaneously at a minimum in the modulation cycle in order to

minimize the (monaural) salience of the instantaneous phase transition (see Figure 1). Stimuli were AM'ed at a rate of 41 Hz. The IPD cycled periodically at three different rates: 1.7 Hz, 3.4 Hz and 6.8 Hz, referred as IPD cycle rates. However, since responses were expected to be elicited by each IPD transition, we refer to the stimulus largely in terms of IPM rate, which indicates the total number of IPD transitions per second, irrespective of direction. Thus, the three IPM rates tested were: 3.4 Hz, 6.8 Hz and 13.7 Hz. These IPM rates corresponded to an IPD transition every 12, 6 and 3 AM cycles, respectively (Figure 1).

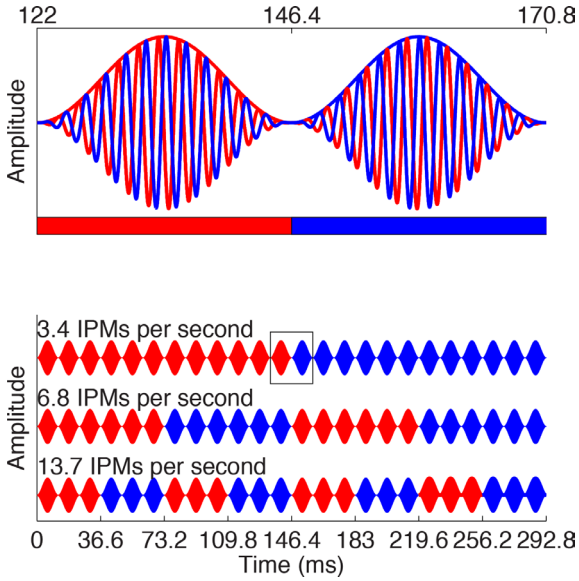


Figure 1. AM stimuli presented to each ear. *Top panel*, red and blue correspond to sounds presented to right and left ears, respectively. Filled horizontal bars indicate the ear with leading phase. The IPD of this example corresponds to ±90° (switching from -45° to 45°, and vice versa, in the two ears). IPD transitions are introduced when the stimulus amplitude is zero. *Bottom panel*, three IPM rates employed. Red regions illustrate an IPD of +90° IPD, whereas blue regions illustrate an IPD of -90° IPD. IPM rate is controlled by the number of AM cycles where the IPD is held constant.

For EEG recordings to ITDs conveyed in the stimulus envelope, 3000-Hz tones were modulated with a transposed envelope (Bernstein and Trahiotis 2002) at 128 Hz, and a second-order AM (41 Hz) applied diotically. IPDs of the 128-Hz envelopes were switched between ±90° IPD (around 0°) or from systematically between 0° and 180° (i.e. transposed phase of -90° in one ear and +90° in the other).

Stimuli were created in Matlab, and presented by an RME Fireface UC sound card (24 bits, 48 kHz sampling rate) connected to Etymotic Research ER-2 insert earphones. Sound level was verified with a 2-cc B&K artificial ear. EEG responses were recorded differentially from surface electrodes; the reference electrode was placed on the vertex (Cz), and the ground electrode on the right clavicle. Two recording electrodes were placed on the left and right mastoid (TP9 and TP10). Electrode impedances were kept below 5 kΩ. Responses were amplified with a 20x gain (RA16LI Tucker-Davis Technologies), and digitalized at a rate of 24.414 kHz, and a resolution of 16 bits/sample (Medusa RA16PA Tucker-Davis Technologies). The cutoff frequencies of the internal bandpass filter were 2.2 kHz and 7.5 kHz, respectively (6 dB per octave). Recordings were next stored on a RX5 Pentusa before being passed to hard disk via custom software. Recordings were processed off-line using Matlab.

During the experiment, subjects sat in a comfortable chair in an acoustically isolated sound booth, and watched a subtitled film of their choice. Subjects were encouraged to sit as still as possible, and were offered a short break every 15-20 minutes. Epochs of each measurement were transformed to the frequency domain (fast Fourier transform (FFT) of 100000 points;

0.24 Hz of resolution) and FFTs from all epochs were averaged. Six frequency bins were tested for significance; corresponding to the IPD cycle rate and the next four harmonics, as well as the frequency bin corresponding to the AM rate.

3 Results

3.1 Sensitivity to IPDs conveyed in the temporal fine structure of low-frequency sounds

EEG recordings were obtained for three IPM rates (3.4 Hz, 6.8 Hz and 13.7 Hz) and seven IPMs (±11.25°, ±22.5°, ±45°, ±67.5°, ±90°, ±112.5° and ±135°), generating a total of 30 conditions (21 dichotic, 9 diotic), each lasting ≈5 min, giving a total recording time of 2.5 h. Figure 2 plots the spectral magnitude of a typical recording for a dichotic condition with IPM of ±67.8° and IPM rate of 6.8 Hz, , in which a significant response was observed for the frequency bin corresponding to the IPM rate (black arrow), and corresponding diotic conditions are shown in Figure 3 - for which no response to the IPM rate was observed (see Ross, 2008). Note that in both conditions the ASSR to the AM rate (41 Hz) was clearly observed.

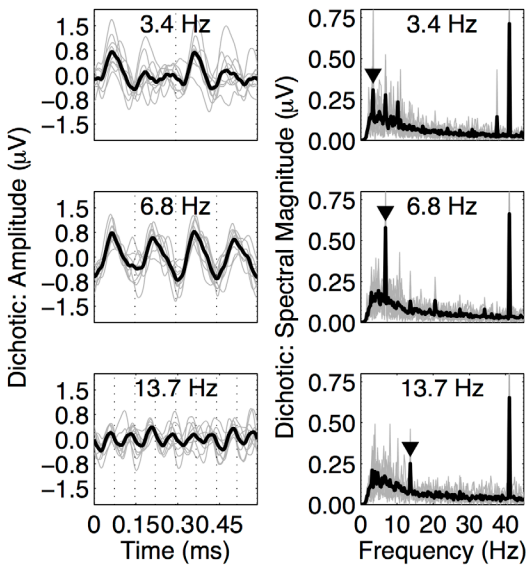


Figure 2. Average (thick line) responses in time and frequency domains to dichotic stimuli, with individual responses shown in thin gray lines. Time domain responses are shown in *left* panels and frequency responses in *right* panels. Responses to low, middle and high IPM rates (3.4 Hz, 6.8 Hz and 13.7 Hz) in *top*, *middle*, and *bottom* rows. All responses correspond to the ±67.8◦ IPD condition. Dashed vertical lines in time-domain responses indicate the time of the IPD transition. Black markers in the frequency domain indicate IPM rate.

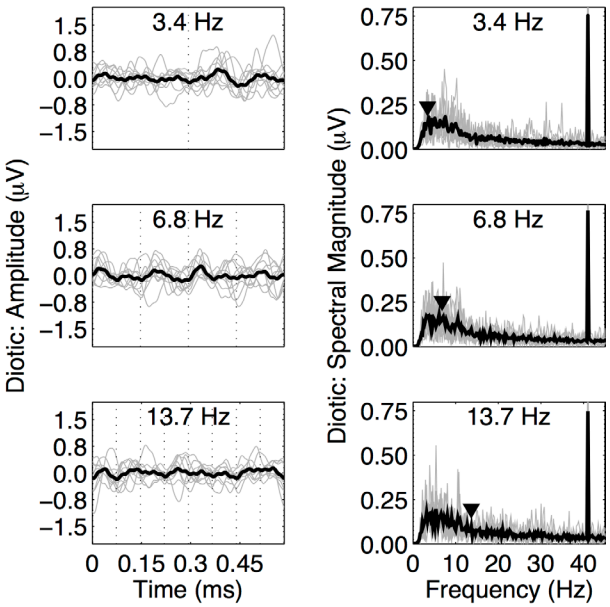


Figure 3. Average (thick line) responses in time and frequency domains to diotic stimuli. Data are plotted in a manner consistent with Figure 2.

Figure 2 shows responses analysed in the time domain (left panels) and the frequency domain (right) for a single IPM condition (±67.8°). Individual (gray lines) and group mean (black) for the slowest IPM rate (3.4 Hz; Figure 3, *top*), typically displayed a P1-N1-P2 response to each IPD transition, with component latencies of approximately 50, 100 and 175 ms, respectively. The next two harmonics (6.8 Hz and 10.2 Hz) were also observed clearly in the response to the 3.4-Hz IPM. Evoked responses to the intermediate IPM rate (6.8 Hz; Figure 2) became steady state at the same rate as the IPM, and so we term this response the IPM-FR. In the time domain (left-middle), the peak amplitude of the IPM-FR is broadly the same as that for the slowest IPM rate (3.2 Hz), but responses were more steady state than those at the low rate, being more consistent in phase and amplitude across epochs (evidenced by a more favourable SNR). This is confirmed by comparing the variability of the individual responses in the time domain as well as the frequency domain (Figure 2), where the spectral magnitude was almost twice larger than that the largest magnitude obtained at the lowest IPM rate. Moreover, the average spectral magnitude of the next harmonics was almost twice smaller than those obtained at the lowest IPM rate. Finally, the IPM-FR is observed at the most rapid IPM rate (13.7 Hz, bottom panels of Figure 3). As for the intermediate IPM rate, responses show a steady-state pattern in the time domain, albeit with a reduced amplitude. Frequency domain analysis revealed the following response occurred primarily at the frequency bin corresponding to the IPM rate. Harmonics were no longer observed in the grand averaged data.

For analysis of data in the frequency domain, Hotelling's T2 tests were applied to the frequency bins corresponding to the IPM rate and the next four harmonics. Responses were classified as significant if either the left (Tp9), right (Tp10), or both electrodes observed a significant response for a given frequency bin. The frequency bin corresponding to the IPM rate (second harmonic) elicited the highest number of significant responses. Indeed, responses obtained at IPM rates of 6.8 Hz and 13.7 Hz were observed for all ten subjects for the ±45° and ±135° IPD conditions, respectively. This was not the case for the lowest IPM rate where significant responses were observed for only seven subjects. In terms of spectral magnitude, responses obtained at the intermediate IPM rate showed the largest magnitude at the second harmonic, i.e. the harmonic corresponding to the IPM rate, consistent with the hypothesis that a following-response (FR) is evoked by IPM.

Finally, we assessed the frequency bin corresponding to the IPM rate. A three-way non-parametric repeated measures ANOVA with factors of electrode position (left or right mastoid),

IPM rate, and IPD revealed that factors IPM rate ($p < 0.001$), IPD ($p < 0.001$) as well as the interaction between IPM rate and IPD ($p = 0.01$) were significant. Responses obtained at 6.8 Hz IPM rate were maximal for IPDs spanning ±45° to ±90°, whereas responses obtained at 13.4 Hz IPM rate increased monotonically with increasing IPD.

3.2 Sensitivity to IPDs conveyed in the temporal envelope of high-frequency sounds

In contrast to low-frequency tones, the magnitude of the IPM-FR for IPDs conveyed in the modulated envelopes of high-frequency (3-kHz) tones was relatively low for IPMs switching between ±90° IPD. Therefore, we also assessed EEG responses to IPMs switching between 0° (in phase) and 180° (anti-phasic) IPD conditions. These latter switches evoked considerably larger EEG responses (Figure 4, *left*). The magnitude of the IPM-FR decreased with increasing carrier frequency, and was sensitive to an offset in the carrier frequency between the ears (Figure 4, *right*).

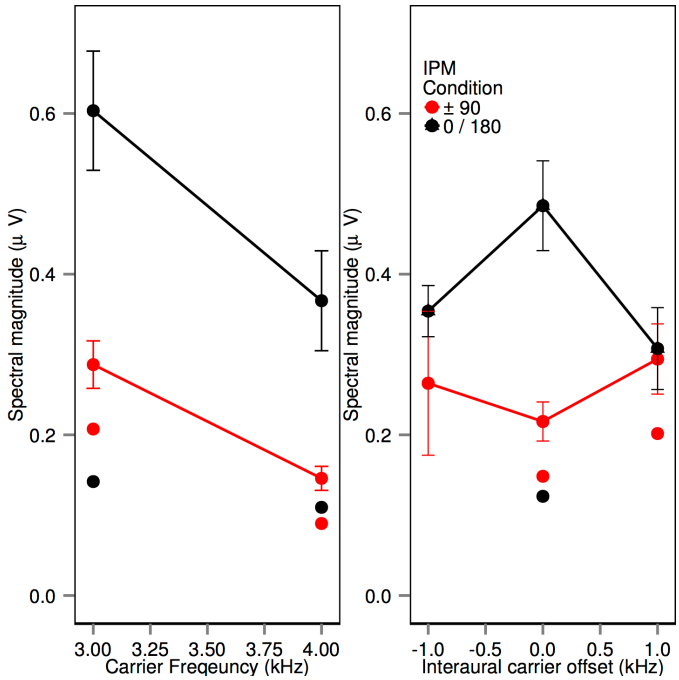


Figure 4. *left* - magnitude of IPM-FRs to 3-kHz transposed tones in which envelope IPDs were switched between ±90°, or between 0° and 180° (transposed phase of 90° at one and -90° at the opposite ear). *right* - magnitude of IPM-FR as a function of carrier frequency mismatch. In both plots, unconnected symbols indicate response magnitude in equivalent diotic control conditions.

4 Discussion

We have demonstrated that an IPM-FR can be reliably obtained and used as an objective measure of binaural processing. Depending on the IPM rate, periodic IPMs can evoke a robust steady-state following response. This is consistent with previous reports that an abrupt change in carrier IPD evokes a P1-N1-P2 transient response (Ross et al., 2007; Ross et al., 2007b; Ross, 2008), and with evidence that abrupt periodic changes in interaural correlation evoked steady-state responses (Dajani and Picton, 2006). The data also demonstrate that the magnitude of the IPM-FR varies with both the IPM rate and the magnitude of the carrier IPD. As diotic carrier phase modulation did not elicit detectable neural responses, the observed responses to IPM must reflect binaural processing rather than the monaural encoding of abrupt phase shifts (Ross, Tremblay, et al., 2007; Ross, Fujioka, et al., 2007; Ross, 2008). The time-domain response waveforms are consistent with evoked responses occurring in cortical and sub-cortical sources

along the auditory pathway, and that they originate from superposition of transient middle-latency responses (Galambos et al., 1981).

The difference in the preferred IPDs for EEG signals conveying IPDs in the low-frequency TFS and the modulated envelopes of high-frequency tones is consistent with the underlying physiological mechanisms of neurons in the brainstem pathways sub-serving these two cues. Low-frequency binaural neurons in the MSO and midbrain (inferior colliculus) are maximally activated by IPDs in the range 45°-90° (see Grothe et al, 2010 for review). This is consistent with their delay sensitivity being generated by largely excitatory inputs from the two ears. Conversely, those in the LSO appear to be excited by one ear, and inhibited by the other (Grothe et al, 2010. Consequently, maximum activity is generated in this (high-frequency) pathway when the sounds at the two ears are out of phase with each other (here, when the signal is switched from 0° to 180°). The IPM-FR, therefore, appears able to distinguish between these two pathways based on the preferred IPDs employed to evoke the response. This provides a possible means of distinguishing which of the binaural pathways is activated in bilateral CI users, as well as a means of objective fitting of devices. Work is currently underway to assess the IPM-FR in bilateral CI users . The EEG recording of steady-state responses in CI users is made challenging by the electrical artefacts created by stimulus pulses, but artefact rejection techniques have been successfully applied for the recording of AM-evoked steady-state responses (e.g., Hofmann and Wouters, 2010). Such techniques should also be applicable for the recording of the IPM-FR.

Acknowledgments

This work is supported by the European Commission under the Health Theme of the 7th Framework Programme for Research and Technological Development.

References

Abel, S. M., Giguère, C., Consoli, A., and Papsin, B. C. (2000). "The effect of aging on horizontal plane sound localization." J. Acoust. Soc. Am. 108, pp. 743–52.

Babkoff, H., Muchnik, C., Ben-David, N., Furst, M., Even-Zohar, S., and Hildesheimer, M. (2002). "Mapping lateralization of click trains in younger and older populations." Hear. Res. 165, pp. 117–27.

Bernstein, L. R. and Trahiotis, C. (2002). "Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli"." J. Acoust. Soc. Am. 122, pp. 1026–1036.

Chait, M., Poeppel, D., de Cheveigné, A., and Simon, J. Z. (2005). "Human auditory cortical processing of changes in interaural correlation." J. Neurosci. 25, pp. 8518–27.

Dajani, H. R. and Picton, T. W. (2006). "Human auditory steady-state responses to changes in interaural correlation." Hear. Res. 219, pp. 85–100.

Galambos, R., Makeig, S., and Talmachoff, P. J. (1981). "A 40-Hz auditory potential recorded from the human scalp." Proc. Natl. Acad. Sci. U. S. A. 78, pp. 2643–7.

Grothe, B., Pecka, M., and McAlpine, D. (2010). "Mechanisms of sound localization in mammals." Physiol. Rev. 90, pp. 983–1012.

Hari, R., Aittoniemi, K., Järvinen, M. L., Katila, T., and Varpula, T. (1980). "Auditory evoked transient and sustained magnetic fields of the human brain. Localization of neural generators." Exp. Brain Res. 40, pp. 237–40.

Hofmann, M., & Wouters, J. (2010). "Electrically evoked auditory steady state responses in cochlear implant users." J. Assoc. Res. Otolaryngol. 11, pp. 267-282.

Liégeois-Chauvel, C., Musolino, a., Badier, J. M., Marquis, P., and Chauvel, P. (1994). "Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components." Electroencephalogr. Clin. Neurophysiol. 92, pp. 204–14.

Lorenzi, C., Gatehouse, S., and Lever, C. (1999). "Sound localization in noise in hearing-impaired listeners." J. Acoust. Soc. Am. 105, pp. 3454–63.

Massoud, S., Aiken, S. J., Newman, A. J., Phillips, D. P., and Bance, M. (2011). "Sensitivity of the human binaural cortical steady state response to interaural level differences." Ear Hear. 32, pp. 114–20.

McEvoy, L. K., Picton, T. W., and Champagne, S. C. (1991). "Effects of stimulus parameters on human evoked potentials to shifts in the lateralization of a noise." Audiology 30, pp. 286–302.

Moore, D. R., Hutchings, M. E., and Meyer, S. E. (1991). "Binaural masking level differences in children with a history of otitis media." Audiology 30, pp. 91–101.

Näätänen, R. and Picton, T. (1987). "The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure." Psychophysiology 24, pp. 375–425.

Picton, T. W., Hillyard, S. A., Krausz, H. I., and Galambos, R. (1974). "Human auditory evoked potentials. I. Evaluation of components." Electroencephalogr. Clin. Neurophysiol. 36, pp. 179–90.

Ross, B. (2008). "A novel type of auditory responses: temporal dynamics of 40-Hz steady-state responses induced by changes in sound localization." J. Neurophysiol. 100, pp. 1265–77.

Ross, B., Fujioka, T., Tremblay, K. L., and Picton, T. W. (2007). "Aging in binaural hearing begins in mid-life: evidence from cortical auditory-evoked responses to changes in interaural phase." J. Neurosci. 27, pp. 11172–8.

Ross, B., Tremblay, K. L., and Picton, T. W. (2007). "Physiological detection of interaural phase differences". J. Acoust. Soc. Am. 121, p. 1017.

Chapter 22

Minimum Audible Angles measured with Simulated Normally-Sized and Oversized Pinnas for Normal-Hearing and Hearing-Impaired Test Subjects

Filip M. Rønne, Søren Laugesen, Niels S. Jensen, Julie H. Pedersen
Eriksholm Research Centre, Snekkersten, Denmark

Contact author: FMR, fmr@eriksholm.com

Abstract

The human pinna introduces spatial acoustic cues in terms of direction-dependent spectral patterns that shape the incoming sound. These cues are specifically useful for localization in the vertical dimension. Pinna cues exist at frequencies above approximately 5 kHz, a frequency range where people with hearing loss typically have their highest hearing thresholds. Since increased thresholds often are accompanied by reduced frequency resolution, there are good reasons to believe that many people with hearing loss are unable to discriminate these subtle spectral pinna-cue details, even if the relevant frequency region is amplified by hearing aids.

One potential solution to this problem is to provide hearing-aid users with artificially enhanced pinna cues – as if they were listening through oversized pinnas. In the present study, it was tested whether test subjects were better at discriminating spectral patterns similar to enlarged-pinna cues. The enlarged-pinna patterns were created by transposing (T) generic normal-sized pinna cues (N) one octave down, or by using the approach (W) suggested by Naylor & Weinrich (2011). The experiment was cast as a determination of simulated minimum audible angle (MAA) in the median sagittal plane. 13 test subjects with sloping hearing loss and 11 normal-hearing test subjects participated. The normal-hearing test subjects showed similar discrimination performance with the T, W, and N-type simulated pinna cues, as expected. However, the results for the hearing-impaired test subjects showed only marginally lower MAAs with the W and T-cues compared to the N-cues, while the overall discrimination thresholds were much higher for the hearing-impaired compared to the normal-hearing test subjects.

Keywords: pinna cues, minimum audible angle, MAA, localization, vertical localization, discrimination, saggital plane, hearing impaired.

1 Introduction

Spatial hearing relies on spatial acoustic cues. The binaural cues (interaural time and level differences, ITD and ILD) are reasonably well preserved with hearing aids, while monaural spectral cues can be substantially changed. The most striking example is that the spectral filtering imposed by the pinna is completely bypassed with behind-the-ear (BTE) hearing aids,

because the microphones sit on top of the pinna. Pinna cues are useful for localisation out of the horizontal plane, and for resolving front-back confusions.

Previous attempts at providing hearing-aid users with natural pinna cues have only been partly successful (Jensen et al., 2013). A likely explanation is that the fine high-frequency spectral details (present from 5 kHz and upwards) of natural pinna cues are at least partly imperceptible by people with hearing impairment, due to reduced frequency selectivity.

Naylor & Weinrich (2011) described a hearing-aid configuration incorporating a method to create directional spectral cues with similar characteristics to pinna cues, but in a lower frequency region around 2-6 kHz, whereby they would be easier to perceive for hearing-impaired test subjects. The hypothesis was that hearing-aid users could learn to utilise these cues and thereby retrieve lost spatial hearing abilities. Hypothesised end-user benefits of such a method could be improved localisation out of the horizontal plane and in the front-back dimension, improved spatial unmasking, and improved spatial awareness in general.

A subsequent question is whether test subjects can learn to exploit a modified set of pinna cues. Hofman et al (1998) and Van Wanrooij & Van Opstal (2005) reported about experiments where test subjects had their pinnae modified by concha moulds. Results show that while localisation performance was disrupted immediately after mould insertion, subjects regainedd their localisation ability over the course of weeks. Typically after 6 weeks of learning they had reached an asymptote close to their previous performance.

1.1 Main Research Question

In brief, the outcome measure of this study's experiment was the Minimum Audible Angle (MAA) in the vertical plane. This is the minimum angular change away from a given reference angle, which can be discriminated in a simulated set-up with sound stimuli presented via headphones. The monaural spatial information is provided by different sets of pinna cues, either simulated normally-sized pinna cues or enhanced pinna cues either transposed one octave down or cues similar to what Naylor & Weinrich, (2011) suggested. Thus, the MAA will be determined with normally-sized (N), transposed (T) or Weinrich (W) pinna cues.

The main research question was: Are the enhanced pinna cues (T or W) easier to discriminate than natural pinna cues (N) for hearing-impaired test subjects? Thus, is the MAA obtained with natural pinna cues (MAA_N) significantly larger than the MAA obtained with the two types of enhanced pinna cues (MAA_T, MAA_W)?

2 Method and Material

2.1 Pinna Cues

Three different cue-sets were used in the experiment, all artificially generated. The normally-sized pinna cues are designed to mimic key trends of recorded pinna cues found in literature (e.g., Shaw et al, 1997). Figure 1 illustrates the response of the normally-sized (N) pinna cue system to different incidence angles in the median saggital plane. Worth noticing is the low-frequency notch that moves upwards as the incidence angle moves away from 0° (0° = straight ahead, 90° = straight up). Figure 2 shows the transposed system (T), which was designed to shift the entire N-cue-structure one octave down. Figure 3 shows the Naylor & Weinrich (2011) cue system (W).

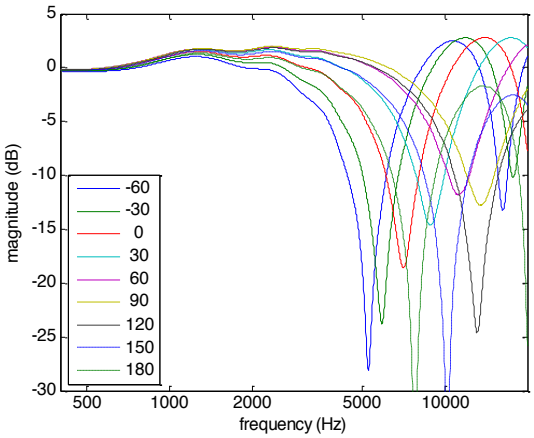


Figure 1. (top left): Response pattern in the median saggital plane for the normal-cue system (N).

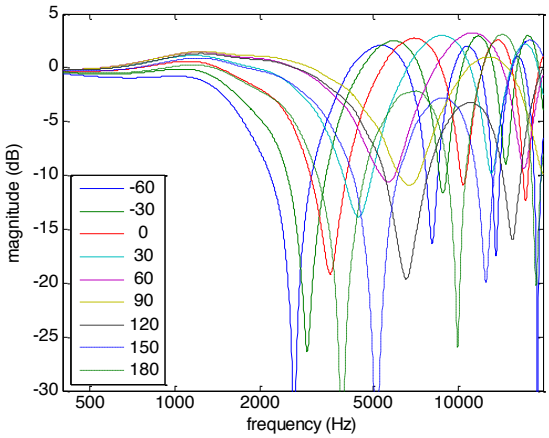


Figure 2. (top right): Response pattern for the transposed-cues system (T).

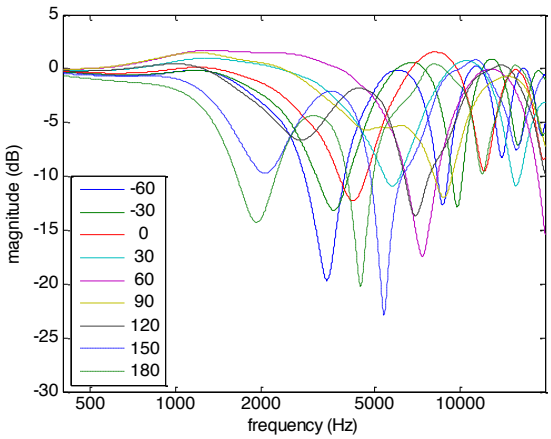


Figure 3. (bottom right): Response pattern for the Naylor & Weinrich, (2011) cue system (W).

2.2 Test Subjects

13 HI test subjects (33 - 79 years old, mean 64 years, SD 15 years) with no history of cognitive problems participated in the study. It was required that the uncomfortable level (UCL) was at least 15 dB higher than the hearing threshold (HTL), to ensure enough headroom for the hearing-loss compensation used to make stimuli audible.

11 test subjects with expected normal hearing were recruited for the NH group (21 - 52 years, mean 39 years, SD 11 years).

All test-subject audiograms were re-measured if older than one year, or if there was a suspicion that changes had occurred. The test subjects' individual hearing threshold levels for the test ear are plotted in Figure 4.

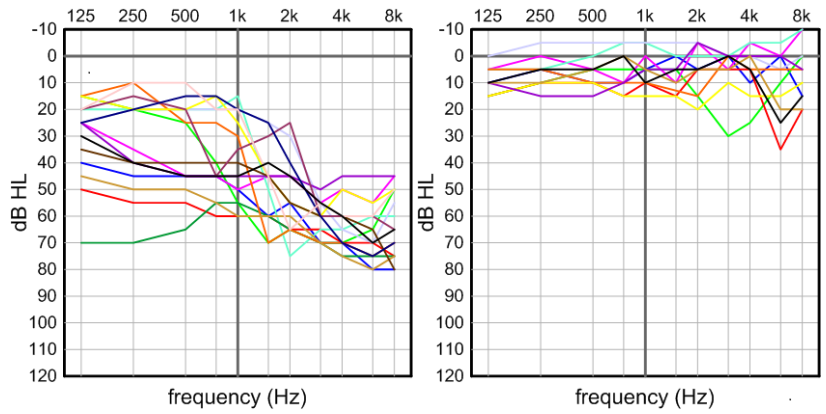


Figure 4. Test-ear hearing thresholds of HI (left) and NH (right) test subjects.

2.3 Psychoacoustic Experiment

An adaptive three-interval, three-alternative forced-choice (3-AFC) discrimination procedure, coupled with a one-up-two-down staircase rule tracking the 70.7%-correct point on the underlying psychometric function, was used.

In each trial, three consecutive stimuli were presented. Two represented the reference incidence angle (0° or 120°) and one represented the variable angle (the 'target') determined by the adaptive procedure. The test subject's task was to determine the diverging stimulus by pressing one of three buttons on a touch screen, on which the target stimulus was randomly assigned to one of the buttons in each trial.

Following each response, the test subject was given visual feedback (about the correct answer) via the screen. In the first trial within a run, the initial incidence angle was easy to discriminate. Pilot testing suggested that appropriate values of the adaptive angle were 60° for the 0° reference and 180° for the 120° reference. The step-size was decreased along the run, with initial step size set to 8°. After each down-up reversal, the step size was halved, until a final step size of 2° was reached. After the final step size was reached, another eight reversals were completed.

The final MAA threshold estimate was calculated based on the median of the last three lower reversal points during a run. On average, this method provides a threshold value, which is lower than that produced by e.g. Levitt (1971) or Neher et al. (2011). However, data inspection indicated that both of these alternate methods could lead to misleading results if a test subject had been performing unreliably during parts of the run. Further, since this study investigates differences between conditions (cues) rather than the absolute MAA values, the consequently lower MAA was accepted.

2.4 Set-Up

The test took place in a single-walled, sound-attenuated test booth. The test software was run on a personal computer, which was connected to a 14" touch screen and a RME Hammerfall DSP Multiface II 24-bit soundcard running at 96 kHz sampling frequency. A Lake People phone-amp G103 was used to drive a pair of Sennheiser HDA 200 headphones.

2.5 Stimuli

All test stimuli were based on white noise samples generated by the test software and steeply band pass filtered to 0.4-16 kHz. The same frozen noise sample was used within each 3AFC triplet, while a new white noise sample was generated for each new triplet. In each triplet, the simulated pinna cues were then added on the fly in Matlab.

Compensation for hearing loss was applied. In short, a level corresponding to 60 dB SPL(C) free-field was selected as the presentation level to be used with NH test subjects. For the HI test subjects, it was then ensured that in each one-third octave band the RMS level of the stimulus was at least 15 dB above hearing threshold before pinna-cue shaping. The required gain was added by designing an IIR filter in the test software to shape the stimulus. Note that the +15-dB gain strategy was terminated at 9 kHz. In higher frequency bands, the same gain as found for 9 kHz was applied, because of the lack of audiogram information. A similar approach was used towards the lowest frequencies.

2.6 Test Protocol

Both the NH and the HI test subjects completed nine test conditions: The MAA test in all combinations of two reference angles (0°/120°) and three cues (N/T/W), and further a front-back discrimination test with all three cues (these results are not reported in this paper).

The order of conditions was counterbalanced for the NH subjects, who completed all nine measurements within one visit. The testing of HI test subjects included two visits. The first visit included audiometry, if the existing audiogram was more than one year old, and the three MAA tests with 120° as reference angle. The second visit included the MAA tests with 0° as reference angle, the front-back tests, and a retest measurement (MAA at 0° with T cues). All visits included the same training conditions, T0 and N120. Within each visit, the order of test conditions was counterbalanced.

The blocking of the 120° measurements introduced a bias (visit effect) in the comparison of the 0° and 120° measurements, but this was preferred over a bias in the comparison between different types of cues, which was the essential comparison in this study.

3 Results

3.1 Data Inspection and Removal

Prior to analysis, the collected data were inspected. The subjective visual inspection of the adaptive MAA measurements pointed out 6 test runs, from which no reliable result could be extracted, as a performance plateau was never reached. These data points were removed.

3.2 Raw Data and Descriptive Statistics

The MAA data for the NH and the HI groups are plotted in Figure 5. The plot show raw and mean (±SD) data for each of the six MAA test conditions.

The main observations which can be made from Figure 5 are:

- Substantially better (lower MAA) performance in the NH group.
- Large individual variation, especially in the HI group.
- Very small deviations between test conditions in the NH group.
- Some deviations between test conditions in the HI group, with better performance with W-cues than with N-cues (with T-cues in between), most pronounced for the 120° conditions.

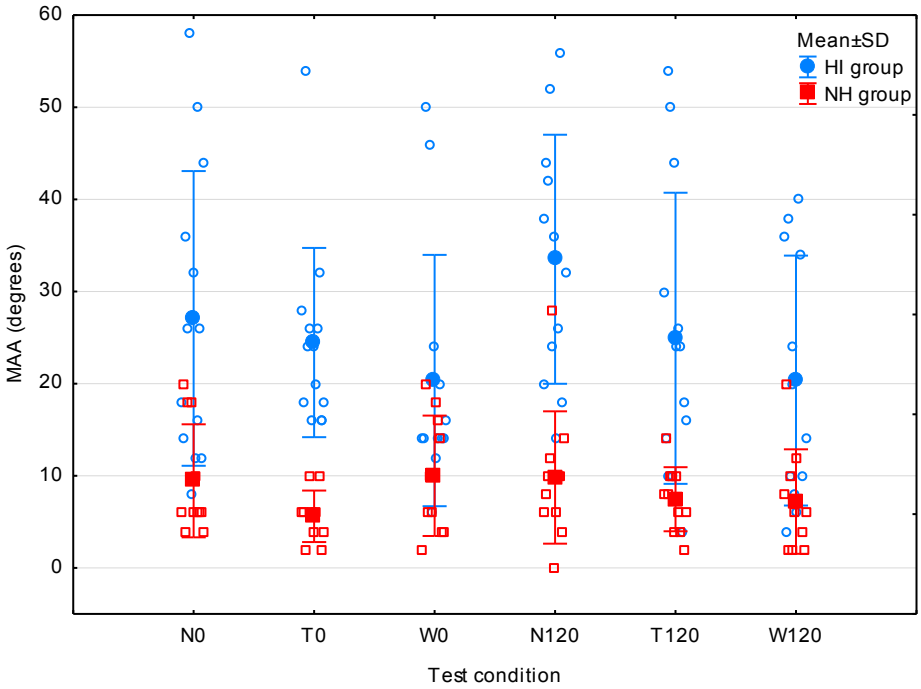


Figure 5. Raw MAA data and mean±SD for the NH and HI groups for each of the six test conditions.

A mixed-model ANOVA was used to analyze the NH MAA data. Besides a significant test-subject effect (p = 0.002) (the random variable "test subject" was only included as a main effect), the various effects in the model came out as being non-significant, as already indicated by the rather similar mean values shown in Figure 5. Thus, the NH-group performance did not depend on either cue type or reference angle.

Another mixed-model ANOVA was used to analyze the HI MAA data. As could be expected based on the large individual variance displayed in Figure 5, a highly significant test-subject effect ($p << 0.001$) was observed. Furthermore, a significant effect of cue type was observed (p = 0.002), while the effect of reference angle was non-significant. No interactions were significant.

Since the difference between normal and enhanced cues was the essence of the research question, the cue-type effect was further explored by conducting two planned comparisons, N vs. T and N vs. W (across the two reference angles). These analyses showed a significant difference between N and W cues (mean MAA difference of around 10°, p = 0.0005) and a non-significant difference between N and T cues (mean MAA difference of around 5°, p = 0.06).

3.3 Training Effects and Test-Retest Variability (Reliability)

Possible training effects, i.e. improved performance over time, were investigated by introducing 'test run number' as a variable. Neither the NH nor the HI subject analysis showed any significant within-session training effects.

The HI tests were conducted during two separate visits, thus enabling a possible between-visit training effect, but with two different sets of test conditions included at the two visits (MAA-120° conditions at first visit, MAA-0° and FB conditions at second visit), it was not possible to assess the between-visit effect. However, the fact that we found no reference-angle effects suggests that the between-visit effect, if present, was limited.

4 Discussion

4.1 Research Question

The hypothesis of the main research question was that the MAA is larger for normal cues than for enhanced cues in a group of HI test subjects. The results showed a significant MAA difference between N and W cues, but not between N and T cues. Thus, the hypothesis was confirmed for the W cues, but not confirmed for the T cues (even though the overall trend was the same in these data).

For the NH group, no differences in performance between cues were observed. This was as expected.

The NH data may be used as a reference in the discussion of the HI data. When the two data sets are compared (see Figure 5), it is noteworthy that even though the use of enhanced cues improves the HI performance in the MAA measurements (from around 30° to around 20°), the mean performance with enhanced cues is still substantially poorer than NH mean performance, which is around 10° with all three types of cues. It may be questioned whether improving the MAA from 30° to 20° will result in any perceivable real-life benefit, but answering this question will obviously require further testing including training test subjects to exploit modified pinna cues.

4.2 Stimuli Level Differences – a Potential Discrimination Cue?

This study investigated whether the frequency-shaping differences caused by changes in pinna-cue filtering across different incidence angles were discriminable by test subjects. However, the subjects did not only have the frequency shaping to listen for, but consequently also a level cue. To quantify whether the level cue was large enough for the test subjects to take advantage of it, the rms-levels were calculated for the three different cue sets. Unfortunately, the specific hearing loss of a specific test subject, and the frequency range one includes in the analysis, very much affects the rms difference between different incidence angles, making it difficult to make conclusions. However, at large angular differences it was clearly observed that the level cue becomes prominent, whereas the level cue decreased with decreasing incidence angle. At average MAAs (< 35° for HI) level-cue magnitudes (< approximately 1dB) were found to be smaller than the level just-noticeable-differences for HI test subjects reported to be 1 to 2dB (Moore, 2003). This indicates that most subjects likely used the frequency-shaping cue rather than the level cue.

The test design could have addressed the potential level-cue issue by applying level roving. However, this would have made the test more difficult to perform, because subjects would have been forced to ignore these added level differences when performing the test. Thus, level roving was rejected to avoid the much more extensive training required, cf. Moore et al. (1989).

5 Conclusions

This study compared the ability to discriminate simulated normally-sized pinna cues with the ability to discriminate two different sets of simulated enhanced pinna cues, in a group of NH test subjects and in a group of HI test subjects, respectively. The results showed that the NH group discriminated equally well with all three sets of pinna cues, showing a mean MAA of around 10° across cue types and reference angle. In the HI group, use of W-cues improved the mean MAA from around 30° to around 20°. This improvement was significant and twice as large as the non-significant mean improvement (from around 25° to 20°) offered by the T-cues.

While the HI group showed an improved ability to discriminate enhanced pinna cues as compared to normal pinna cues, the improvements were not as large as expected, and performance was still substantially poorer than that observed in the NH group.

References

Hofman, P.M., Van Riswick, J.G., Van Opstal, A.J., 1998. Relearning sound localization with new ears. *Nat. Neurosci.*, 1(5), p.417–421.

Jensen, N.S., Neher, T., Laugesen, S., Johannesson, R.B., Kragelund, L., 2013. Laboratory and Field Study of the Potential Benefits of Pinna Cue-Preserving Hearing Aids. *Trends Amplif.*, 17(3), p.171–188.

Levitt, H., 1971. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49, p.467–477.

Moore, B.C., 2003. *An introduction to the psychology of hearing*, Academic press San Diego.

Moore, B.C., Oldfield, S.R., Dooley, G.J., 1989. Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *J. Acoust. Soc. Am.*, 85, p.820–836.

Naylor, G., Weinrich, S.G., 2011. System and method for generating auditory spatial cues. United States Patent, US 7,936,890, May 3, 2011

Neher, T., Laugesen, S., Jensen, N.S., Kragelund, L., 2011. Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *J. Acoust. Soc. Am.*, 130(3), p.1542–1558.

Shaw, E., Anderson, T., Gilkey, R., 1997. Binaural and Spatial Hearing in Real and Virtual Environments, chapter Acoustical features of human ear. *RH Gilkey TR Anderson Lawrence Erlbaum Assoc. Mahwah NJ USA*, p.25–47.

Van Wanrooij, M.M., Van Opstal, A.J., 2005. Relearning sound localization with a new ear. *J. Neurosci.*, 25(22), p.5413–5424.

Chapter 23

Moving Objects in the Barn Owl's Auditory World

Ulrike Langemann, Bianca Krumm, Katharina Liebner, Rainer Beutelmann, Georg M. Klump
Hearing4all, Oldenburg, Germany

Contact author: GMK, georg.klump@uni-oldenburg.de

Abstract

Barn owls are keen hunters of moving prey. They have evolved an auditory system with impressive anatomical and physiological specializations for localizing their prey. Here we present behavioural data on the owl's sensitivity for discriminating acoustic motion direction in azimuth that, for the first time, allow a direct comparison of neuronal and perceptual sensitivity for acoustic motion in the same model species. We trained two birds to report a change in motion direction within a series of repeating wideband noise stimuli. For any trial the starting point, motion direction, velocity (53-2400°/s), duration (30-225 ms) and angular range (12-72°) of the noise sweeps were randomized. Each test stimulus had a motion direction being opposite to that of the reference stimuli. Stimuli were presented in the frontal or the lateral auditory space. The angular extent of the motion had a large effect on the owl's discrimination sensitivity allowing a better discrimination for a larger angular range of the motion. In contrast, stimulus velocity or stimulus duration had a smaller, although significant effect. Overall there was no difference in the owls' behavioural performance between "inward" noise sweeps (moving from lateral to frontal) compared to "outward" noise sweeps (moving from frontal to lateral). The owls did, however, respond more often to stimuli with changing motion direction in the frontal compared to the lateral space. The results of the behavioural experiments are discussed in relation to the neuronal representation of motion cues in the barn owl auditory midbrain.

Keywords: auditory motion discrimination, motion-direction sensitivity, sound localization, bird

1 Introduction

Our auditory world is by no means stationary. Natural sound sources often change position and also listeners move in their environment. Predators like the barn owl will routinely hunt moving prey, and they perform in-flight corrections (Hausmann et al., 2008). Thus, with respect to sound source localization, real world situations confront the auditory system with varying binaural cues. Possible motion-related dynamic binaural cues are the change in interaural time differences (ITDs), in interaural level differences (ILDs) or in spectral peaks and notches. Generally, ITD cues identify sound source azimuth, ILD cues azimuth and/or elevation and peaks and notches mostly elevation (e.g., Grothe et al., 2010). It has been suggested that auditory motion processing relies on specialized circuits composed of directionally tuned and motion sensitive neurons in the midbrain (reviewed by Wagner et al., 1997). Time constants for processing of non-stationary dynamic cues must be short to allow tracking of moving sound sources (Wagner, 1991). However, time constants for binaural processing have found to depend on measuring procedures (e.g., Shackleton & Palmer, 2010).

Auditory motion perception has been investigated simulating motion by switching between neighbouring free-field loudspeakers (e.g., Wagner & Takahashi, 1992) or by varying interaural cues in headphone presentation (reviewed by Middlebrooks & Green, 1991). Neurophysiological correlates to human auditory motion perception have been investigated in animal models with both free-field and headphone stimulation (e.g., McAlpine et al., 2000; Wagner et al., 1997). What is missing, however, is the direct comparison of neuronal and perceptual sensitivity for acoustic motion in the same model species.

Here we report auditory motion perception in the barn owl, a species for which data on the neuronal representation of motion stimuli in the inferior colliculus (IC) are available (e.g., Wager & Takahashi, 1992; Kautz & Wagner, 1998; Wagner & von Campenhausen, 2002; Wang & Pena, 2013). Specifically, Wagner and Takahashi (1992) showed that the angular velocity and stimulus duration affected the neurons' response. Here, we investigate how stimulus velocity, the size of the angle of sound incidence (i.e., the angular range), and stimulus duration affect the owls' perceptual performance. Furthermore, we analyse whether the perception of changes in motion direction differs for stimuli presented in the frontal or in the lateral space. Finally, we will discuss the behavioural data with reference to the responses of motion direction sensitive neurons in the barn owl auditory midbrain.

2 Methods

We trained two barn owls to report a change in auditory motion within a series of repeating band-pass noise stimuli (500-6000Hz) presented in the horizontal plane. Motion was simulated by sequentially activating up to eight loudspeakers (Vifa XT25TG30-04), fading in and out the sound between adjacent loudspeakers (spacing 12°; Fig. 1). The experiments were performed in a sound-attenuating echo-reduced chamber (IAC 1203-A, walls with sound absorbing foam). Signals were generated by a 24-bit sound card (Hammerfall DSP Multiface II, RME, Germany) and the loudspeakers were driven by a multichannel amplifier (RMB-1048, Rotel, Japan). The barn owls had learned to sit on a waiting perch during the repeated reference stimuli and to fly to a report perch when a test stimulus was presented (Go/NoGo paradigm). Within a trial the reference stimuli (repetition period 1300ms) had the same direction, while the test was a single stimulus presented with opposite motion direction. Each stimulus started at a random position between -42° and +42° in azimuth. This prevented the owls from using location *per se* rather than motion direction to solve the task. The owls were also trained to generally orient their head in the waiting position towards 0° in azimuth. Correct responses were rewarded. The waiting intervals before a test stimulus were randomized. For any trial the direction of the noise sweeps, the velocity (53-2400°/s) and the effective duration (30-225 ms, with 10 ms Hanning ramps) and angular range of the motion (12-72°) of the reference stimuli were randomly selected from a distribution of combinations of parameters. Each combination of parameters was presented 20 times. Experimental sessions with 44 trials each (36 test trials and 8 catch trials having no change in motion direction) were included in the analysis if the owls responded at least ten times (10 "Hits").

We performed two experimental series. Series A had effective stimulus durations of 75, 150 and 225 ms and provided stimulus velocities of between 53 and 960°/s. Series B had effective stimulus durations of 30, 60 and 90 ms and provided stimulus velocities of between 133 and 2400°/s. Note that stimulus duration and stimulus velocity are interdependent variables. The relation between sensitivity and the parameters stimulus duration, angular range, and stimulus velocity was analysed by linear regression. We will use the regressions' coefficient of determination (R^2) and associated ANOVA results to describe the contribution of each of the three parameters to the perception of a change in motion direction.

We made the following assumptions for data analysis. (1) Noise sweeps moving from the owl's lateral space toward the frontal space were classified as moving "inward", noise sweeps moving

from the frontal space toward the lateral space were classified as moving "outward". (2) "Lateral" stimuli were noise sweeps that stayed in one hemisphere more the 50% of the time, but were allowed to cross the midline up to 15° to the contralateral side, "frontal" stimuli were noise sweeps moving more the 50% of the time in the range between plus and minus 15° from the midline (Fig. 1). The reasoning for this approach is that the owls' space maps in the left and right IC overlap by about this amount (Knudsen & Konishi, 1978a,b). We limited our analysis to noise sweeps of an extent of 36° and 48° to avoid floor and ceiling effects. We then compared the owl's performance for inward and outward stimuli and for stimuli between the frontal and lateral space using Chi^2 tests.

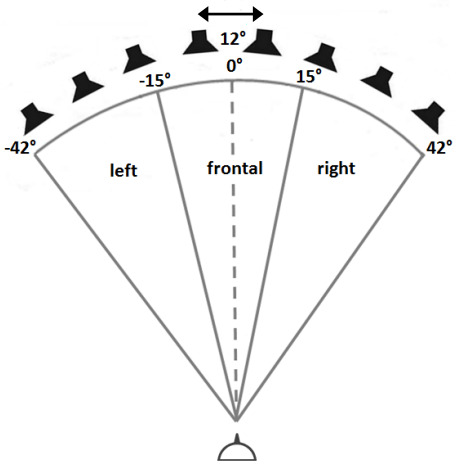


Figure 1. Stimuli were classified as moving either "inward" (from lateral to frontal) or "outward" (from frontal to lateral). "Frontal" stimuli were moving at least 50% of the time in the range between plus and minus 15° from the midline. "Lateral" stimuli were moving more than 50% of the time in one of the two hemispheres (including up to 15° of the other hemisphere).

3 Results

In a behavioural task, barn owls show high sensitivity for auditory motion discrimination. Sensitivity was reduced only for small angular ranges. Furthermore auditory motion discrimination was still good at the shortest stimulus duration (30ms) and at high velocities (up to 2400°/s).

3.1 Effects of Stimulus Velocity, Size of the Angular Range, and Stimulus Duration on Auditory Motion Discrimination

Our data show that the owls' sensitivity for auditory motion discrimination increased with increasing angular range. The amount of variance accounted for (R^2) by "angular range" was between 0.70 and 0.80 for each of the conditions shown in Figures 2 and 3. Stimulus duration and stimulus velocity had only a small, though significant effect on sensitivity. When comparing sensitivity at similar angular ranges, sensitivity moderately increased for longer stimuli and decreased for faster moving stimuli (R^2 values for "duration" were between 0.03 and 0.13, R^2 values for "velocity" were between 0.07 and 0.21, all $P \leq 0.001$). The same pattern was represented by the ANOVA results: the factor "angular range" had F-values of at least 854, F-values for "duration" were between 12 and 55, F-values for "velocity" were between 25 and 95 (all $P \leq 0.001$). The owls' sensitivity was still rather high, if stimulus duration was further decreased in series B (Fig. 3). When comparing sensitivity for noise sweeps of similar velocities at different durations (Fig. 2, 3), we had expected to find them to be similar across conditions. In contrast to our expectation, stimulus velocity was a rather bad predictor of the owls' behavioural performance. This can be explained, however, by the interaction of duration and angular range that determines stimulus velocity.

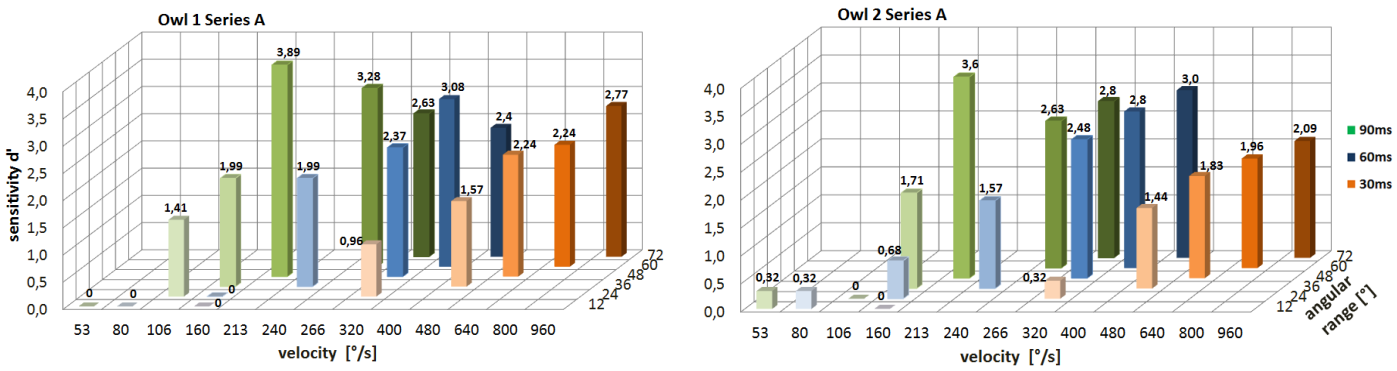


Figure 2. Series A with stimulus velocities up to 960°/s: Sensitivity increases with longer stimulus duration and larger angular range. Sensitivity index (d') is plotted as a function of stimulus velocity, angular range and duration.

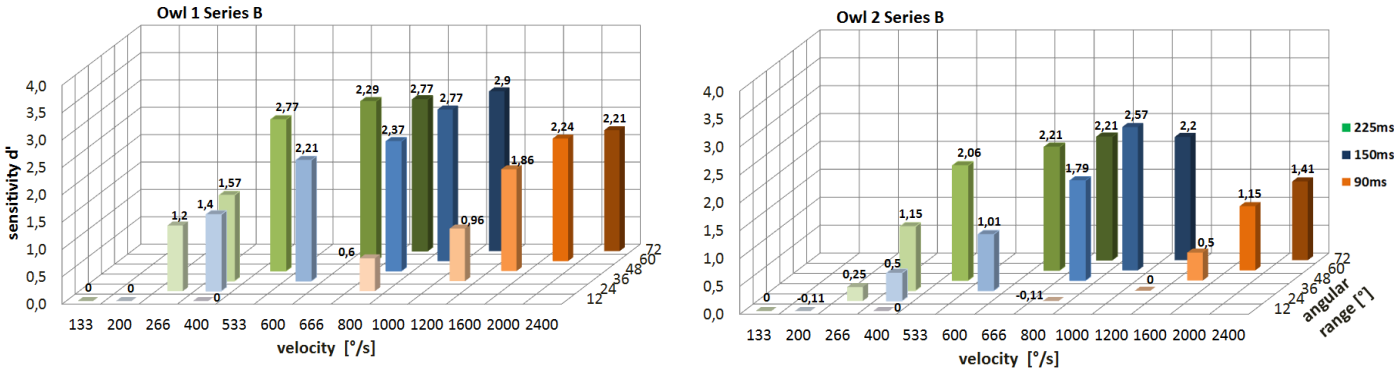


Figure 3. Series B with stimulus velocities up to 2400°/s: Sensitivity is hardly affected when stimulus velocity is further increased. Sensitivity index (d') is plotted as a function of stimulus velocity, angular range and duration.

3.2 Effects of Stimulus Position on Auditory Motion Discrimination

Responses to inward and outward sweeps were equally frequent: the owls had relative hit rates of between about 35 and 43 % for each of the two directions (Chi^2 tests n.s., Table 1, left panel). The owls responded more often to stimuli with a change in direction moving in the frontal range compared to those moving in the lateral range (Table 1, right panel): Owl 1 had a relative Hit rate of about 67% in the frontal space and of only 38% in the lateral space ($Chi^2 = 15.7$, $P < 0.01$) while owl 2 had relative hit rates of about 52% and 39% for the frontal and lateral range, respectively ($Chi^2 = 3.1$, $P = 0.076$.).

Table 1. The absolute "Hit" and "Miss" rates combining Series A and B. Left panel: Test stimuli were moving either inward or outward. Right panel: Noise sweeps were presented from lateral or frontal.

Owl 1	inward	outward	
Hit	29	28	$Chi^2 = 0{,}009$ $p = 0{,}926$ $N = 158$
Miss	52	49	
Owl 2			
Hit	28	38	$Chi^2 = 0{,}862$ $p = 0{,}353$ $N = 170$
Miss	53	51	

Owl 1	lateral	frontal	
Hit	51	52	$Chi^2 = 15{,}743$ $p = < 0{,}001$ $N = 214$
Miss	85	26	
Owl 2			
Hit	66	38	$Chi^2 = 3{,}131$ $p = 0{,}077$ $N = 243$
Miss	104	35	

4 Discussion

Using simulated acoustic motion stimuli, we have demonstrated how the sensitivity for discriminating auditory motion direction in the horizontal plane is related to stimulus velocity, stimulus duration, and the size of the angular range covered by the moving sound. Furthermore, we describe how motion discrimination is affected by the location in the owl's auditory space. Here we relate these data to results from studies of motion-direction sensitive neurons in the barn owl.

Motion direction sensitive neurons have been observed in the barn owl IC and optic tectum (Wagner et al. 1997, Wagner & von Campenhausen 2002). Neurophysiological studies on directional sensitivity in auditory motion processing by the owl have applied a wide range of angular velocities. Witten et al. (2006) presented simulated motion based on ITD change at a speed of 32°/s observing shifts of auditory spatial receptive fields in relation to motion direction that could be explained by adaptation. Similarly, Wang and Peña (2013) showed that adaptation affected direction selectivity for simulated motion of 470°/s in free-field stimulation. Wagner and Takahashi (1992) observed a broad tuning of motion direction selective units with respect to stimulus velocity when presenting simulated motion with an angular velocity ranging between 120°/s and 1200°/s. They suggest that inhibition accounts for the directional sensitivity. Our behavioural data obtained with noise sweeps in the free field and with an angular velocity between 53°/s and 2400°/s are in accord with the observed broad angular velocity tuning. Stimulus velocity had a moderate effect on motion direction sensitivity provided that the stimulus duration was not too short.

Duration of the noise sweeps per se had only a moderate effect on the behavioural sensitivity. Sensitivity was lower for short than for long stimulus durations. We could observe the lowest sensitivity for stimuli of 30 ms duration. This result is in accord with data describing the barn owl's binaural temporal window for ITD processing ranging from 4 to 42 ms (Wagner, 1991). Inhibitory neuronal processing has been observed to be effective already between 20 and 40 ms after stimulus onset (Wagner, 1990). Inhibition could thus contribute to the dynamic sharpening of the ITD representation in the barn owl's IC and to the duration of the binaural temporal window.

The angular range covered by the noise sweeps in our experiments exerted the strongest effect on the owl's sensitivity. To solve the task, the angular range had to be at least 24°. The angular range necessary for motion direction discrimination was much larger than the mean error for locating points in azimuth as indicated by the head orienting response towards the sound source elicited by broadband noise bursts (Knudsen et al., 1979) or by the owl's minimum audible angle (Bala et al., 2007). Using a pupillary dilation response, Bala and colleagues (2007) measured a

minimum audible angle of 3° in azimuth. In the horizontal plain, Knudsen et al. (1979) observed a localization error being as low as 2° for broadband sounds in the frontal direction (open-loop conditions). The azimuth localization error increased to 6° if the target location was 50° lateral. The small localization error in comparison to the minimum angular range needed for motion direction discrimination suggests that at least partially different neuronal circuits are involved in the location of stationary and of moving sound sources. Wagner and Takahashi (1992) suggested a motion detector circuit in the barn owl auditory midbrain that included inhibition and a delayed interaction between neurons representing different locations in space (Kautz & Wagner, 1998; Wagner et al., 1997). This circuit is interwoven with the computational auditory space map in the barn owl midbrain and relies on the space map providing input (Wagner et al., 1997). Thus, the barn owl brain uses separate but intermingled systems for representing auditory motion and spatial location. Furthermore, the large difference between the localization accuracy and the minimum angular range needed for motion direction discrimination renders it unlikely that the owls solve the motion direction discrimination task by a "snapshot" mechanism. The snapshot hypothesis suggests that the processing of motion simply relies on comparing starting and endpoint of the moving stimulus (e.g., Perrot & Marlborough, 1989). The snapshot hypothesis was also dismissed in human psychophysical studies determining the minimum audible movement angle (reviewed by Middlebrooks & Green, 1991). The difference between the barn owl's sensitivity for comparing angular locations of static sounds (Bala et al., 2007) and the large angular range needed for direction discrimination (present study) suggests that the snapshot hypothesis does not apply to barn owls.

Neurophysiological recordings in the barn owl midbrain have revealed an interaction between the starting point of the motion and the preferred direction of motion direction sensitive units (Wagner & von Campenhausen, 2002; Wang & Peña, 2013). However, the results from the two studies were contradictory. Wagner and von Campenhausen (2002) observed that neurons responded preferentially to outward motion whereas Wang and Peña (2013) observed a preference to inward motion. The behavioural data obtained in the present study are not suited to resolve the discrepancy. Our owls were equally sensitive in detecting inward and outward moving noise sweeps.

Previous barn owl studies indicated that the frontal space is represented by a larger population of neurons than the lateral space (e.g., as reflected by the distribution of best ITDs in the IC; Wagner et al., 2007). Furthermore, the space maps in each IC represent the contralateral space plus an angular range of about 15° from the midline in the ipsilateral space (Knudsen & Konishi, 1978a,b). This results in an overlap in the representation of the frontal space (-15° to 15°) by IC neurons from both sides of the brain. Finally, neurons in the IC tuned to locations in the frontal space have more confined spatial receptive fields (Knudsen & Konishi, 1978b). The neurophysiological results have prompted Knudsen and Konishi (1978b) to predict that localization accuracy should be larger in the frontal space than in the lateral space. As expected, Knudsen et al. (1979) observed a smaller localization error in the frontal field than in the lateral field when measuring the barn owls' head orienting response towards a stationary sound source. Our results are consistent with these findings. The barn owls' motion direction discrimination was more sensitive for stimuli being presented in the frontal space compared to stimuli presented in the lateral space.

Acknowledgments

This study was funded by a grant from the Deutsche Forschungsgemeinschaft (TRR 31).

References

Bala, A.D., Spitzer, M.W., & Takahashi, T.T. (2007). Auditory spatial acuity approximates the resolving power of space-specific neurons. *PLoS One*, 2(8), e675.

Grothe, B., Pecka, M., & McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiological Reviews*, 90, 983-1012.

Hausmann, L., Plachta, D.T.T., Singheise, M., Brill, S., & Wagner, H. (2008). In-flight corrections in free-flying barn owls (*Tyto alba*) during sound localization tasks. *Journal of Experimental Biology*, 211, 2976-2988.

Kautz, D., & Wagner, H. (1998). GABAergic inhibition influences auditory motion-direction sensitivity in barn owls. *Journal of Neurophysiology,* 80, 172-185.

Knudsen, E.I., & Konishi, M. (1978a). A neural map of auditory space in the owl. *Science*, 200, 795-797.

Knudsen, E.I., & Konishi, M. (1978b). Space and frequency are represented separately in auditory midbrain of the owl. *Journal of Neurophysiology*, 41(4),870-884

Knudsen, E.I., Blasdel, G.G., & Konishi, M. (1979). Sound localization by the barn owl (*Tyto alba*) measured with the seach coil technique. *Journal of Comparative Physiology A*, 133, 1-11.

McAlpine, D., Jiang, D., Shackleton, T.M., & Palmer, A.R. (2000). Responses of neurons in the inferior colliculus to dynamic interaural phase cues: evidence for a mechanism of binaural adaptation. *Journal of Neurophysiology*, 83, 1356–1365.

Middlebrooks, J.C, & Green, D.M. (1991). Sound localization by human listeners. *Annual Reviews of Psychology*, 42, 135-159.

Perrott, D.R., & Marlborough, K. (1989). Minimum audible movement angle: marking the end points of the path traveled by a moving sound source. *Journal of the Acoustical Society of America*, 85(4), 1773-1775.

Shackleton, T.M., & Palmer, A.R. (2010). The time course of binaural masking in the Inferior colliculus of Guinea pig does not account for binaural sluggishness. *Journal of Neurophysiology,* 104, 189-199

Wagner, H. (1990). Receptive fields of neurons in the owl's auditory brainstem change dynamically. *European Journal of Neuroscience*, 2, 949-959.

Wagner, H. (1991). A temporal window for lateralization of interaural time difference by barn owls. *Journal of Comparative Physiology A,* 169, 281-289.

Wagner, H., & Takahashi, T. (1992). Influence of temporal cues on acoustic motion-direction sensitivity of auditory neurons in the owl. *Journal of Neurophysiology,* 68(6), 2063-2076.

Wagner, H., & von Campenhausen, M. (2002). Distribution of auditory motion-direction sensitive neurons in the barn owl's midbrain. *Journal of Comparative Physiology A,* 188, 705-713.

Wagner, H., Kautz, D., & Poganiatz, I. (1997). Principles of acoustic motion detection in animals and man. *Trends in Neurosciences,* 20(12), 583-588.

Wagner, H., Asadollahi, A., Bremen, P., Endler, F., Vonderschen, K., & von Campenhausen, M. (2007). Distribution of interaural time difference in the barn owl's inferior colliculus in the low- and high-frequency ranges. *Journal of Neuroscience*, 27(15), 4191-4200.

Wang, Y., & Pena, J.L. (2013). Direction selectivity mediated by adaption in the owl's inferior colliculus. *The Journal of Neuroscience*, 33(49), 19167-19175.

Witten, I.B., Bergan, J.F., & Knudsen, E.I. (2006). Dynamic shifts in the owl's auditory space map predict moving sound location. *Nature Neuroscience,* 9(11), 1439-1445.

Chapter 24

Change Detection in Auditory Textures

Yves Boubenec[1,2], Jennifer Lawlor[1,2], Shihab Shamma[1,2,3], Bernhard Englitz[1,2,4]

[1] *Laboratoire des Systèmes Perceptifs, CNRS UMR 8248, Paris, France*
[2] *Département d'études cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France*
[3] *Neural Systems Laboratory, University of Maryland in College Park, MD, USA*
[4] *Department of Neurophysiology, Donders Centre for Neuroscience, Radboud Universiteit Nijmegen, Netherlands*

Contact author: YB, boubenec@ens.fr

Abstract

Many natural sounds have spectrotemporal signatures only on a statistical level, e.g. wind, fire or rain. While their local structure is highly variable, the spectrotemporal statistics of these auditory textures can be used for recognition. This suggests the existence of a neural representation of these statistics. To explore their encoding, we investigated the detectability of changes in the spectral statistics in relation to the properties of the change.

To achieve precise parameter control, we designed a minimal sound texture — a modified cloud of tones — which retains the central property of auditory textures: solely statistical predictability. Listeners had to rapidly detect a change in the frequency marginal probability of the tone cloud occurring at a random time.

The size of change as well as the time available to sample the original statistics were found to correlate positively with hit rate and negatively with reaction time, suggesting the accumulation of noisy evidence. In summary we quantified dynamic aspects of change detection in statistically defined contexts, and found evidence of integration of statistical information.

1 Introduction

For many natural sounds, such as wind, fire, rain, water or bubbling, integration of statistical information under continuous conditions is critical to assess the predictability of changes. While their spectrogram is dominated by small, recognizable elements (e.g. gust, crackles, drops), their occurrence is not predictable *exactly*, but they occur randomly in time and frequency, yet constrained by certain probabilities. As shown previously, these probabilities/statistics define the identity of such sounds, so called auditory textures (McDermott et al., 2013). McDermott and others could demonstrate this point directly, by recreating auditory textures from samples, using only the spectro-temporal statistics of those sounds (e.g. McDermott & Simoncelli, 2011).

Detecting relevant changes in auditory textures is hence complicated due to the lack of deterministic spectrotemporal predictability, and instead relies on the prediction of stimulus statistics (and their dynamics). The aim of this study is to determine how changes in stimulus statistics are detected under continuous presentation of stimulus. Concretely, we assessed how first-order non-uniform sound statistics are integrated and how listeners disentangle a change in these from the intrinsic stochastic variations of the stimulus.

We designed a broadband stimulus, composed of tone-pips, whose occurrence is only constrained by a frequency marginal probability. While the central 'textural' property of solely

statistical predictability is maintained, it is devoid of most other systematic properties, such as frequency modulation, across channel or temporal correlations. A change is introduced by modifying the marginal distribution at a random time during the stimulus presentation, which subjects are asked to report.

This change detection task allows us to explicitly address the integration of statistical information under continuous conditions, which models more closely the real-world challenge of detecting changes in complex ongoing auditory scenes. In this context, we studied the dependence of detection on the time and size of the change. We found evidence for integration of statistical information to predict a change in stimulus statistics, both based on hit rate as well as on reaction times. The main predictors appear to be the time available for integration of evidence, both within and across trials.

2 Methods

2.1 Participants

Fifteen subjects (mean age 24.8, 8 females, higher education: undergraduate and above) participated in the main study in return for monetary compensation. All subjects reported normal hearing, and no history of psychiatric disorders.

2.2 Experimental Setup

Subjects were seated in front of a screen with access to a response box in sound-attenuated booth (Industrial Acoustics Company GmbH). Acoustic stimulus presentation and behavioural control were performed using custom software package written in MATLAB (BAPHY, NSL, University of Maryland). The acoustic stimulus was sampled at 100 kHz, and converted to an analog signal using an IO board (National Instruments NI PCIe-6353) before being sent to diotic presentation using headphones (Sennheiser HD380, calibrated flat, i.e. ±5dB within 100Hz-20000Hz). Reaction times were measured via a custom-built response box and collected by the same IO card with a sampling rate of 1 kHz.

2.3 Stimulus Design

We used a simplified sound texture model, which retained the property of being predictable only from the statistical properties of its complex spectro-temporal structure (Fig. 1A). The texture was a tone cloud consisting of a sequence of 30-ms, temporally overlapping pure tones whose frequency covered a range of 2.2 octaves (400 to 1840 Hz) divided in 8 frequency bins. The frequency resolution of the tone distribution was 12 semitones per octave, starting at 400 Hz. This allowed 26 tone frequencies in total with 3-4 frequency values in each frequency bin.

The minimal temporal unit of the stimulus was a 30-ms chord in which the number of tones in a particular frequency range was drawn according to a marginal distribution for that frequency range. On average, for each chord duration the mean number of tones per octave was 2. The marginal distribution of occurrence probability was obtained by modifying a uniform distribution in each frequency bin. The probability in each of these is called $P_{uniform}$=⅛=0.125. To generate different initial frequency marginals, we perturbed the marginals randomly by adding/subtracting fixed value Δ corresponding to 50% of the probability (which would be observed for a uniform distribution ($\Delta=P_{uniform}/2=0.0625$)). The resulting marginal distribution was thus pseudo-random with 3 bins at $P_{uniform}+\Delta$, 3 bins at $P_{uniform}-\Delta$ and 2 bins left intact at $P_{uniform}$. This implies that the average initial probability in 2 bins can take 5 different values, namely Δ, $3\Delta/2$, 2Δ, $5\Delta/2$, and 3Δ.

The change consisted in an increment of the marginal distribution in the selected frequency bins at a random point in time (referred to as *change time*) during stimulus presentation. We chose to

use an increment of the marginal distribution in two adjacent frequency bins on the basis that an appearance is more salient than its opposite in a complex acoustic stimulus (Constantino et al., 2012). After the change, the second stimulus continued for up to 2 s or until the subject made an earlier decision, whichever happened first.

The increment size, referred to as *change size*, was drawn from a set of discrete values [50, 80, 110, 140]%, relative to the single bin probability in a uniform distribution (for 8 bins, the single bin uniform probability is 1/8th, and thus a 50% change size would be 1/16th). In order to maintain the overall number of tones per chord before and after the change, the occurrence probability in the six remaining frequency bins was decreased accordingly.

The time at which the change occurred (*change time*) was drawn randomly from an exponential distribution (mean: 3.2 s) limited to the interval [0, 8] s. This choice of distribution prevents subjects from developing a timing strategy, as the instantaneous probability of a change in the next time step is constant.

2.4 Procedure

The experiment was separated into three phases: instruction, training, and main experiment. After reading the instructions, subjects went through 10 min of training (60 trials), where they were required to obtain at least 40% performance. The training comprised only stimuli of the two greatest change sizes (110%, 140%). Three subjects did not attain the criterion level of performance and were not tested further.

The main experiment was composed of two sessions of about 70 min each, comprising a total of 930 trials, corresponding to 60 repetitions of each condition. The two sessions were never more than two days apart. After reading the instructions, subjects were aware that the change could arise at any moment on each trial and that their task was to detect it within the 2 s window.

Visual feedback was always displayed on the screen in front of them, either a red square was displayed when the button was pressed before the change (false alarm), or when the button was not pressed within the 2 s time window after the change (miss). A green square was displayed when the button was pressed after the change but within the 2 s window.

In addition, sound level was roved from trial to trial ; it was chosen randomly between 60 and 80 dB SPL (sound pressure level). This procedure is classically applied to prevent subjects from adopting any absolute level strategy. The inter-trial interval was ∼1 s with a small, random jitter (<0.1 s) depending on computer load.

2.5 Data Analysis

We quantified the ability of the subjects to detect the change in stimulus statistics using two measures, hit rate and d-prime (d'). We also found reaction times to depend on the difficulty.

These measures were computed as a function of change size and change time. Since change times were distributed continuously but with an exponential distribution, the set of change times was binned with approximately exponentially increasing bin size (in order to achieve comparable numbers of trials in each bin).

To control for inattentive subjects, we set a 35% threshold for the total false alarm rate. Two subjects were discarded according to this criterion leaving a total of 10 subjects for the data analysis, with false alarm rates around 25%.

Hit rate and reaction times

We computed a subject's hit rate as the fraction between successful detection (hits) out of the total trials for which the change occurred before the subject's response (hits + misses). False alarms were excluded from the hit rate computation, since they occurred before the subject was

exposed to the change (see d' below for an inclusion of false alarms). We obtained reaction times by subtracting the change time from the response time in each trial.

d' Analysis

We computed d' values to assess the ability to detect changes, while taking their false alarm rate into account. Due to the present task structure, d' was computed as a function of time from stimulus onset (see Fig. 1E for an illustration, approximated as d'(t)=Z(HR(t))-Z(FAR(t)), where Z(p) is the inverse of the cumulative Gaussian distribution. HR(t) is the hit rate as a function of time t since stimulus onset. Hit rate was computed as the fraction of correct change detections, in relation to the number of trials with changes occurring at t. Similarly, the false alarm rate FAR(t) was computed as the number of false alarms that occurred over all 2 s windows (starting at t), in which no change in statistics occurred. The window of 2 s was chosen to be compatible with the hit rates in the 2 s decision window. d' was computed separately for different times and change sizes, yielding only a limited number of trials per condition. To avoid degenerate cases (i.e. d' would be infinite for perfect scores), the analysis was not performed separately by subject, but over the pooled data. Confidence bounds (95%) were then estimated by grouping data from all subjects. The analysis was verified on surrogate data from a random responder (binomial with very low p at each point in time), providing d' very close to 0 on a comparable number of trials.

Hit rate Dynamics

In order to compare the performance dynamics for different change sizes, we fitted (least-square non-linear minimization) a cumulative Erlang distribution to the data according to:

$$P(\Delta_c, t_c) = P_0(\Delta_c) + P_{max} \quad (\Delta_c) * \gamma(k, t_c/\tau(\Delta_c))/(k-1)!$$

where P_0 is the minimal hit-rate, P_{max} is the maximal hit rate, t_c is change time, $\Delta_c$the change size, γ the incomplete gamma function, τ the function rate, and k controls the function shape. k was kept constant across subjects and change sizes, assuming the shape of the hit rate curves is invariant, which appeared to be the case in our sample.

2.6 Statistical Analysis

In the statistical analysis only non-parametric tests were used. One-way analysis of variance were computed with Kruskal-Wallis' test, the two-way were computed using Friedman's test. Unless in indicated otherwise, error bars correspond to twice the standard error of the mean (SEM). All statistical analyses were performed using Matlab (The Mathworks, Natick).

3 Results

Human listeners (n=15, 10 performed to criterion level) were presented with a continuous acoustic texture and required to detect a change in stimulus statistics, which could occur at a random time, after which they had 2s to respond. Changes occurred in the probability of occurrence of tones in randomly chosen spectral regions (Fig. 1A).

Several parameters of the change could determine its salience, such as its size, timing. These parameters were tested in a single experiment with ~1000 trials over two sessions per subject.

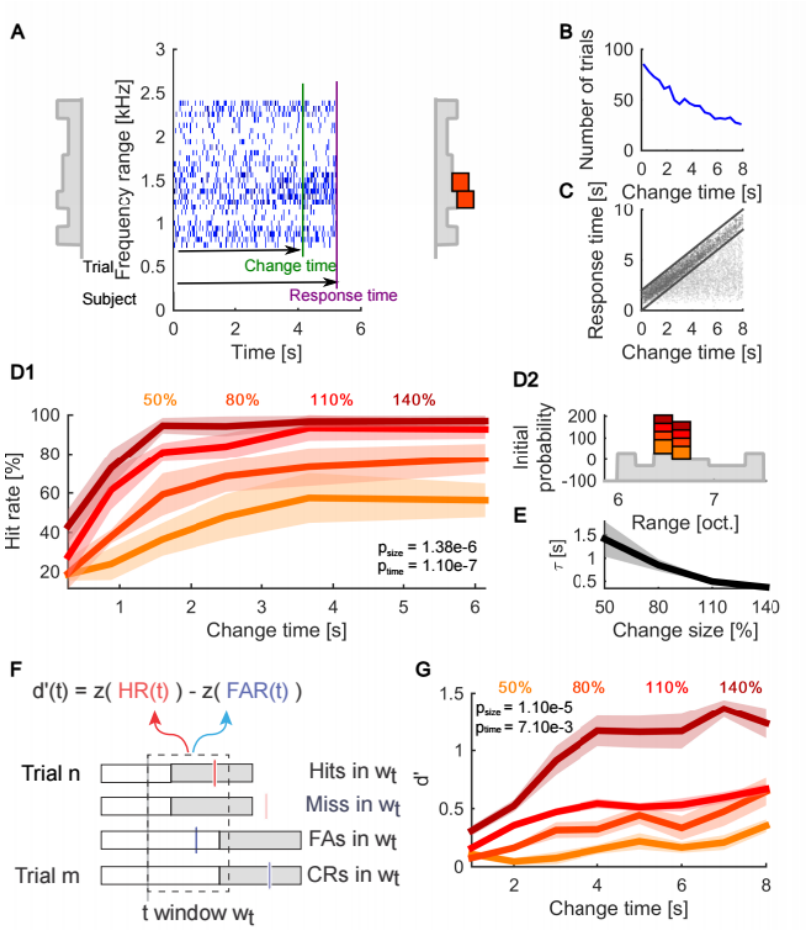


Figure 1. Detection of change in statistics requires integration over time. **(A)** Subjects listened to an acoustic textural stimulus, whose sole predictability was governed by its marginal frequency distribution (grey curve, left). Tones in individual frequency bins were all drawn independent consistent with the marginal (middle). Listeners were instructed to report changes by a button press. The frequency marginal was modified after a randomly chosen point in time ('change time', red bars on the right). The probability in two frequency bins separated by a random distance were modified at a time. All other bins were slightly reduced in probability to maintain the overall loudness constant. **(B)** Distribution of change times across all subjects (n=10). Change times were drawn from an exponential distribution. **(C)** Response times occurred before (false alarms) and after the change time (hits). Subjects usually responded only after an initial time of listening, allowing them to acquire the acoustic statistics. **(D)** Hit rate of change detection depended significantly on change time (x-axis) and change size (colour). **(E)** Four different change sizes were used (colour-coded). **(F)** The dynamics of the hit rate curve also varied with change size, indicated by the fitted parameter τ of a cumulative Erlang distribution (see *Methods*). **(G)** To control for false alarms, d' was also computed as a function of time into the trial (see *Methods* for details), where within each trial the statistics change where the gray area starts.. **(H)** Similar to hit rate, d' increased significantly over time into the trial, and with change size. Errorbars indicate confidence bounds two SEM based on a bootstrap across subjects.

3.1 Detection of Changes in Statistics is Consistent with Integration

The ability of subjects to detect a change in stimulus statistics (Fig. 1A) improved with the time they were allowed to listen to the initial statistics of the texture (Fig. 1D, progression along ordinate, performance measured as hit rate, $p<10^{-6}$, Friedman. Hit rate increased monotonically to an asymptotic level for all change sizes (4 levels, [50, 80, 110, 140]%). Asymptotic hit rate depended on the change size, with bigger changes in marginal probability leading to greater

asymptotic hit rate (Fig. 1D, different colours). Although asymptotic hit rate was above chance (26.6%) for all change sizes, the increase with change size was large and significant (from 50% to 95%, $p<10^{-7}$, Friedman test).

Change size also influenced the shape of the dependence on change time, such that greater change sizes lead to improved hit rate already for shorter change times (Fig. 1D). This translates to a combined steepening and leftward shift of the hit rate curves with change size. Significance of this effect was assessed by fitting sigmoidal hit rate curves of individual subjects with a parametric function (see *Methods*) in order to extract the change size-dependent time constant (Fig. 1F). Hit rate time constants τ significantly decreased with respect to change sizes ($p<10^{-6}$, Friedman test).

The results above stay true if false alarms are taken into account (Fig. 1H). For this purpose, we computed a time-dependent d' measure (see Fig. 1G for an illustration, and *Methods*). Classical d' analysis does not apply in the present dynamic decision paradigm, since the eventual outcome of a trial is not accessible at early times (i.e. when computing d'(0.5 s), a trial with a hit at 5 s should not be counted as a hit, but as a 'correct reject up to 0.5 s; but when computing d'(4 s), the same trial will be counted as a hit). In accordance with the perceived difficulty of the detection task, d' values are significantly positive (Fig. 1H, based on the 95% confidence bounds) and increase with change time (Friedman test, $p<0.01$), but stay relatively low especially for 110% and below, indicates how challenging the task was, leading to a substantial number of false alarms.

The above results are consistent with predictions from statistical integration of information. First, unsurprisingly, change detection should improve with greater change sizes. Second, change detection should improve, but saturate, as the estimation of the statistics of the first stimulus converges, i.e. if longer observation time is available. Third, change detection of bigger changes should require less precision in the estimation, translating to higher hit rate already for shorter observation times.

3.2 Reaction Times are Consistent with Integration

The dependence of hit rate on change time suggests a dynamical integration mechanism to play a role in the decision process. While the dependence on change time indicates an ongoing estimation of the initial statistics, it does not provide much insight into the estimation of the post-change statistics. We hypothesized that reaction times could be here be informative. Reaction times had previously been demonstrated to depend on task difficulty (Kiani et al., 2014), and difficulty in the present paradigm correlates inversely with the availability of a greater difference in change size, i.e. statistics (Fig. 2A).

Reaction time distributions changed both in time and shape as a function of change size (Fig. 2B). Median reaction time correlated negatively with change sizes ($p<0.001$; Fig. 2C), in a manner that corresponded to the inverse of the increase in hit rate with larger change sizes. In addition, the onset time of the distributions decreased with change size, together suggesting that certainty was reached earlier for bigger step-sizes.

As a function of change time, the reaction time distribution changed in a qualitatively similar manner, however, less pronounced (Fig. 2E). Median reaction times correlated negatively with change times, mirroring dependence of hit rate on change times ($p<0.001$; Fig. 2F). This dependence can already be seen in the raw data (Fig. 1C), where hit trials (black) for longer change times exhibit shorter reaction times. Again the onset time decreased correspondingly with longer change time, suggesting more accurate estimation of the initial statistics.

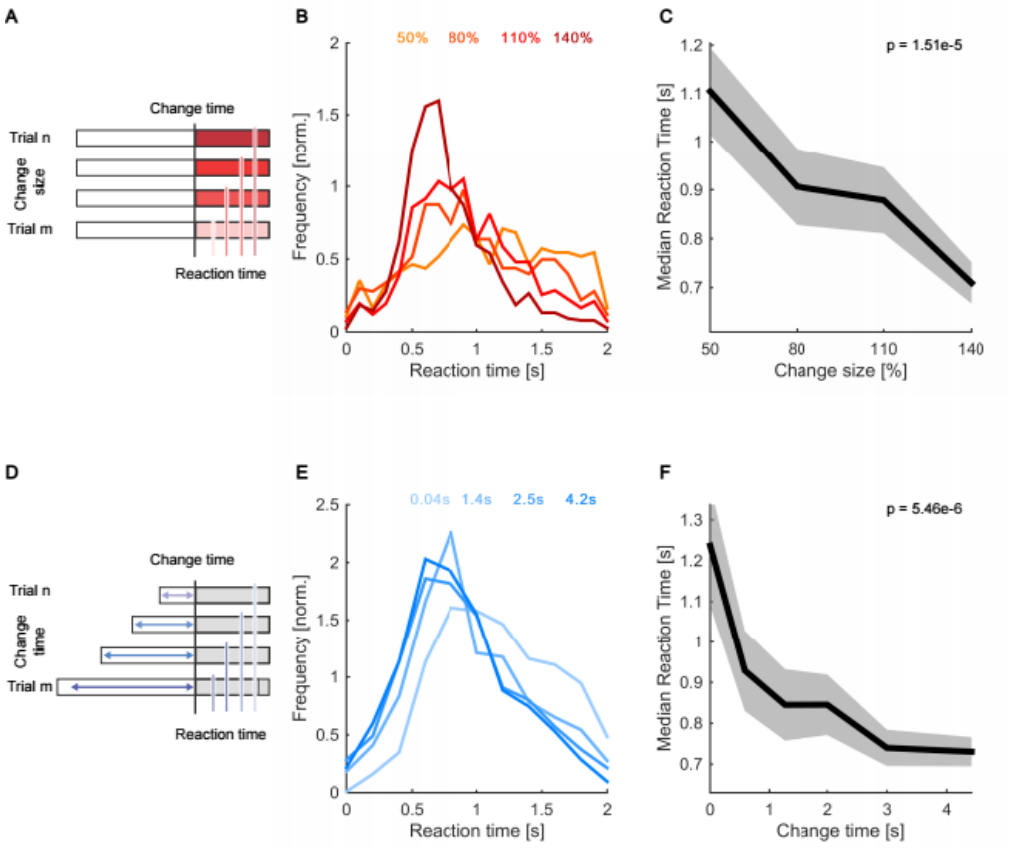


Figure 2. Reaction times reflect task difficulty and temporal integration. **(A-C)** If reaction times are separated by change size **(A)**, one observes a progression in the shape **(B)** as well as median response time **(C)**. The shape of the reaction time distribution sharpens with increasing step-size and the median response time decreases significantly (p=1.51e-05, Kruskal-Wallis). **(D-F)** If reaction times are separated by the duration of the change time before the change in statistics occurred **(D)**, the distribution of reaction times again sharpens with longer change times **(E)** and the median reaction time **(F)** decreases significantly (p=5.46e-06, Kruskal-Wallis).

4 Discussion

In the present study we used a simplified acoustic texture to study detecting changes in the statistics of complex acoustic stimuli. The results suggest that humans are able to integrate stimulus statistics, as expressed in their hit rate and reaction times depending on the parameters of the task.

4.1 Dynamic Representation of Spectral Statistics

The stimulus of the present paradigm can be considered a stochastic version of the appearance or increase in loudness of a sound source within a natural auditory scene. Its design is a compromise between complexity of the spectrotemporal structure and simplicity in specification since the parameter-space is reduced to a minimum and parameters are straightforward to interpret.

The present paradigm builds on research from the spectral and the temporal domain, e.g. profile analysis (Green & Berg, 1991) or some paradigms in mismatch negativity. For example in profile analysis subjects assess the spectral shape of stimuli. In this context it has also been demonstrated that subjects can compare spectra, however, the results were not interpretable in a statistical sense, due to several reasons. First, the time scale of presentation is much longer in

the present paradigm than typically in profile analysis, where the stimulus is presented for multiple seconds rather than subsecond presentations. Second, in the present context the stimulus is composed of randomly timed occurrences of tones in certain frequency bins (which defines the spectral profile over time), rather than static profiles from constantly presented tones or filtered noise. Third, the comparison between different profiles are performed dynamically, rather than in a two stimulus comparison task. As described above this allowed us to study the dynamic integration of stochastic information, and enables the collection of reaction times as another indicator of the underlying processes.

4.2 Future directions

The present study lays some groundwork for future studies, by investigating some of the basic, previously unaddressed questions related to the integration of statistics in complex auditory stimuli, such as auditory textures. Further studies are required to address variations of the present paradigm, such as reductions in probability and continuous stimulus presentation. Attention was not controlled for in the present setting, although subjects were instructed to attend to the stimulus. Some subject reported improved detection of change, when *not* attending to the stimulus, suggesting that a dual-task paradigm would be interesting where attention can be assessed and controlled.

Acknowledgments

Funding was provided through the Advanced ERC ADAM 295603, the ANR-10-LABX-0087-IEC and the ANR-10-IDEX-0001-02-PSL*.

References

Constantino, F. C., Pinggera, L., Paranamana, S., Kashino, M., & Chait, M. (2012). Detection of Appearing and Disappearing Objects in Complex Acoustic Scenes. *PLoS ONE*, doi:10.1371/journal.pone.0046167

Green, D. M., & Berg, B. G. (1991). Spectral Weights and the Profile Bowl. *The Quarterly Journal of Experimental Psychology Section A*, 43(3), 449-58.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron* 84(6): 1329-1342.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron* 71(5): 926-40.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary Statistics in Auditory Perception. *Nature Neuroscience* 16(4): 493-98.

Chapter 25

The Relative Contributions of Temporal Envelope and Fine Structure in Mandarin Lexical Tone Perception in Auditory Neuropathy Spectrum Disorder

Shuo Wang, Ruijuan Dong, Dongxin Liu, Luo Zhang

Otolaryngology - Head & Neck Surgery, Beijing Tongren Hospital, Beijing Institute of Otolaryngology, Capital Medical University, Beijing, China

Li Xu

Communication Sciences and Disorders, Ohio University, Athens, Ohio, USA

Contact author: SW, shannonwsh@aliyun.com

Abstract

Previous studies have demonstrated that temporal envelope (E) is sufficient for speech perception, while fine structure (FS) is important for pitch perception and lexical tone perception for normal-hearing (NH) listeners. Listeners with sensorineural hearing loss (SNHL) have an impaired ability to use FS in lexical tone perception due to the reduced frequency resolution. Listeners with auditory neuropathy spectrum disorder (ANSD) may have deficits in temporal resolution. Little is known about how such deficits may impact their ability to use E and FS to perceive lexical tone, and whether it is the deficit in temporal resolution or frequency resolution that may lead to more detrimental effects on FS processing in pitch perception. Three experiments were conducted in the present study to address the above-mentioned questions. Experiment I used the "auditory chimera" technique to investigate how listeners with SNHL and with ANSD would achieve lexical tone recognition using either the E or the FS cues. Experiment II tested the frequency resolution for the two groups of listeners as measured with the $Q_{10\,dB}$ values of the psychophysical tuning curves (PTCs). Experiment III tested their temporal resolution as measured with the temporal gap detection (TGD) threshold. The results showed that the listeners with SNHL had reduced frequency selectivity, but intact temporal resolution ability, while the listeners with ANSD had degraded temporal resolution ability, but intact frequency selectivity. In comparison with the SNHL listeners, the ANSD listeners had severely degraded ability to process the FS cues and thus their ability to perceive lexical tone mainly depended on the ability to use the E cues. These results suggested that, in comparison with the detrimental impact of the reduced frequency selectivity, the impaired temporal resolution may lead to more degraded FS processing in pitch perception.

1 Introduction

Acoustic signals can be decomposed into temporal envelope (E) and fine structure (FS). Previous studies have demonstrated that E is sufficient for speech perception (Shannon et al., 1995;

Apoux et al., 2013), while FS is important for pitch perception (Smith et al., 2002) and lexical tone perception (Xu & Pfingst, 2003). Mandarin Chinese is a tone language with four phonologically distinctive tones that are characterized by syllable-level fundamental frequency (F0) contour patterns. These pitch contours are described as high-level (tone 1), rising (tone 2), falling-rising (tone 3), and falling (tone 4). It has been demonstrated that FS is also important for lexical tone perception in listeners with sensorineural hearing loss (SNHL), but as their hearing loss becomes more severe, their ability to use FS becomes degraded so that their lexical tone perception relies increasingly on E rather than FS cues (Wang et al., 2011). The reduced frequency selectivity may underlie the impaired ability to process FS information for SNHL listeners. Auditory neuropathy spectrum disorder (ANSD) is an auditory disorder characterized by dys-synchrony of the auditory nerve firing but normal cochlear amplification function. Several studies have demonstrated that ANSD listeners have a dramatically impaired ability for processing temporal information (Zeng et al., 1999; Narne, 2013). It is important to assess the ability of lexical tone perception for ANSD listeners and to investigate how they use FS and E cues to perceive lexical tone, as these results may shed light on the effects of frequency resolution and temporal resolution on FS processing in pitch perception.

2 Lexical tone perception

Experiment I was aimed to investigate the ability of lexical tone perception for both SNHL and ANSD listeners, and to assess the relative contributions of FS and E to their lexical tone perception using the "auditory chimera" technique (Smith et al., 2002; Xu & Pfingst, 2003). As described in Wang et al. (2011), the chimeric tokens were generated in a condition with 16 channels. These 16 FIR band-pass filters with nearly rectangular response were equally spaced on a cochlear frequency map, ranging from 80 to 8820 Hz. In order to avoid the blurriness between E and FS in the chimeric stimuli, we adopted a lowpass filter (cut-off at 64 Hz) for extraction of the envelopes from the filters. By analysing the tone responses to the chimeric tone tokens, we could understand the competing roles of FS and E in tone perception in various groups of listeners.

2.1 Methods

Three groups of adult subjects, including 15 NH subjects (8 females and 7 males), 16 patients with SNHL (10 females and 6 males), and 27 patients with ANSD (9 females and 18 males), participated in the present study. The subjects with SNHL had hearing loss ranging from moderate to severe degree, and the ANSD subjects had hearing loss ranging from mild to severe. The original tone materials were consisted of 10 sets of Chinese monosyllables with four tone patterns for each, resulting in a total of 40 commonly used Chinese words. These original words were recorded using an adult male and an adult female native Mandarin speaker. The tone tokens in which the durations of the four tones in each monosyllable were within 5-ms precision were chosen as the original tone tokens. These tone tokens were processed through a 16-channel chimerizer in which the FS and E of any two different tone tokens of the same syllable were swapped. For example, two tokens of the same syllable but with different tone patterns were passed through 16 band-pass filters to split each sound into 16 channels. The output of each filter was then divided into its E and FS using a Hilbert transform. Then, the E of the output in each filter band was exchanged with the E in that band for the other token to produce the single-band chimeric wave. The single-band chimeric waves were summed across all channels to generate two chimeric stimuli. Finally, a total of 320 tokens were used in the tone perception test, including 80 original unprocessed tone tokens and 240 chimeric tone tokens. A four-alternative, forced-choice procedure was used for the tone perception test.

2.2 Results

Fig. 1 (left panel), plots the accuracy of tone perception to the original tone tokens for the NH, SNHL, and ANSD subjects. The median tone perception performance was 97.2%, 86.5%, and 62.8% correct, respectively for the three groups. The performance slightly reduced for the subjects with SNHL, while the ANSD subjects had dramatic decreases in their tone perception performance also with a very large individual variability. The right panel of Fig. 1 plots the mean percentages and standard deviations of tone perception responses to the chimeric tone tokens. The subjects with NH relied on FS nearly entirely to perceive lexical tones regardless of E. In comparison with the NH subjects, the subjects with SNHL depended more on E cues to perceive tones, as they had impaired ability to use FS. In contrast to the finding in the NH and SNHL subjects, only 26.5% of the tone responses were consistent with the FS of the chimeric tone tokens for the ANSD subjects, indicating that the ANSD subjects had more severely impaired ability to use FS in tone perception. On the other hand, the ANSD subjects relied more heavily on E cues in perceiving lexical tones, indicating that the ability of the ANSD subjects to use E cues for lexical tone perception still remained at a reasonable level.

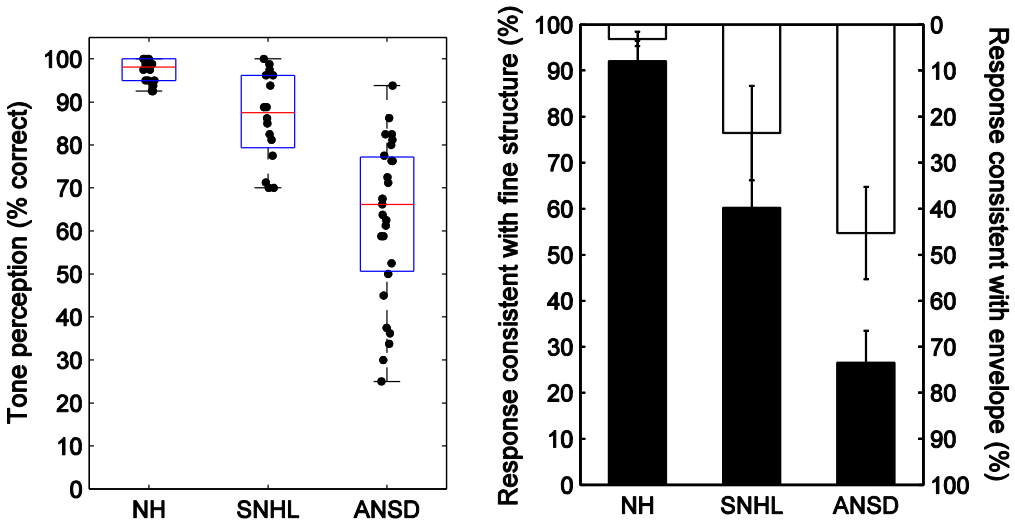


Figure 1. Tone perception performance. Left: Boxplots show the tone perception performance with the original, unprocessed tone tokens for the subjects with NH, SNHL, and ANSD. The box represents the 25th to 75th percentiles and the horizontal line in the box represents the median. The whiskers represent the range of the data. Right: Mean and standard deviation of tone perception with the chimeric tone tokens for the subjects with NH, SNHL, and ANSD. Black bars represent responses consistent with FS (right ordinate) and open bars represent responses consistent with E (left ordinate).

3 The frequency resolution and temporal resolution

Studies have shown that listeners with SNHL may have reduced frequency resolution (Kluk & Moore, 2006), but probably normal temporal resolution (Glasberg et al., 1987). In contrast, listeners with ANSD may have a dramatically impaired ability for processing temporal information (Narne, 2013), while their frequency selectivity may be close to normal (Vinay & Moore, 2007). To clarify whether it is the deficit in temporal resolution or frequency resolution that may lead to more detrimental effects on FS processing in lexical tone perception, we also measured the frequency resolution using the mean $Q_{10\,dB}$ values of the psychophysical tuning curves (PTCs), and the temporal resolution using the temporal gap detection (TGD) threshold for the SNHL and ANSD subjects.

3.1 Methods

Three groups of subjects who were recruited in Experiment I also participated in Experiment II in which we measured the psychophysical tuning curves (PTCs) of the auditory system using a fast method called SWPTC, developed by Sek & Moore (2011). The subject was required to detect a sinusoidal signal, which was pulsed on and off repeatedly in the presence of a continuous noise masker. The PTCs at 500 and 1000 Hz in each ear were tested separately. The sinusoidal signals were presented at 15 dB sensation level (SL). The masker was a narrowband noise, slowly swept in frequency. The frequency at the tip of the PTC was estimated using a four-point moving average method (4-PMA), and the values of $Q_{10\,dB}$ (i.e., signal frequency divided by the PTC bandwidth at the point 10 dB above the minimum level) was used to assess the sharpness of the PTC. The greater the $Q_{10\,dB}$ values, the sharper the PTC.

Experiment III tested the temporal gap detection (TGD) threshold for all subjects using a TGD program developed by Zeng et al. (1999). The test stimuli were generated using a broadband (from 20 to 14,000 Hz) white noise. The noise had a duration of 500 ms with 2.5-ms cosine-squared ramps. A silent gap was produced in the center of the target noise. The other two reference noises were uninterrupted. The TGD threshold of left and right ear was measured separately for each subject using a three-alternative, forced-choice procedure. The intensity level of the white noise was set at the most comfortable loudness level for all subjects.

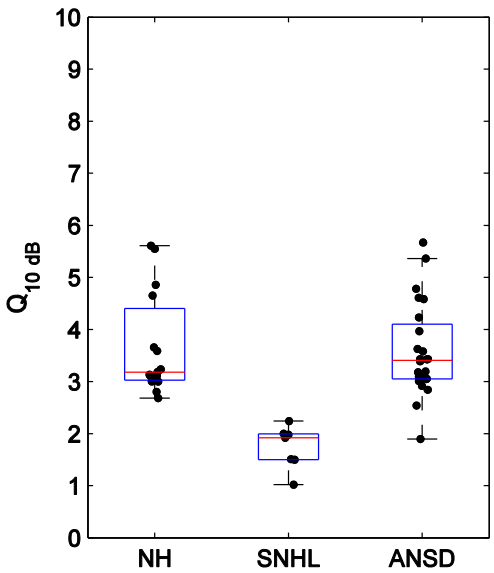


Figure 2. The $Q_{10\,dB}$ values of psychophysical tuning curves for subjects with NH, with SNHL, and with ANSD. The box represents the 25th to 75th percentiles and the horizontal line in the box represents the median. The whiskers represent the range of the data. Data points outside of the whiskers are outliers. Note that for 9 of the 16 subjects with SNHL, the PTCs were too broad so that the Q_{10dB} values could not be determined.

3.2 Results

The values of $Q_{10\,dB}$ were determined in all 15 NH subjects, 7 of the 16 SNHL subjects, and 24 of the 27 subjects with ANSD. Note that for 9 of the 16 subjects with SNHL, the PTCs were too broad so that the $Q_{10\,dB}$ values could not be determined. Since no significant differences were found between the $Q_{10\,dB}$ values at 500 and 1000 Hz, the $Q_{10\,dB}$ values averaged across the two test frequencies are plotted in Fig. 3. Statistically significant differences of the $Q_{10\,dB}$ values were found between the subjects with NH and those with SNHL and between the subjects with ANSD and those with SNHL, while no statistically significant differences were present between the subjects with NH and those with ANSD. This indicates that a majority of the subjects with ANSD

showed close-to-normal frequency resolution, whereas the SNHL subjects had poor frequency resolution in the present study.

Figure 3 shows the averaged TGD thresholds of both ears for the three groups of subjects. Welch's test in a one-way ANOVA was found that significant difference of TGD thresholds among the three groups of subjects [$F(2,56)=9.9$, $p < 0.001$]. Post hoc Tamhane' T2 correction test analysis indicated that the mean TGD thresholds for both ears was significantly higher in the listeners with ANSD (11.9 ms) than in the listeners with SNHL (4.0 ms; $p < 0.001$) and in the NH (3.9 ms; $p < 0.001$) listeners; while the subjects with SNHL had the TGD thresholds close to normal. It is notable that the variability in TGD thresholds for the ANSD subjects was very large, ranging from normal limits to 10 times norm. This result indicates that the SNHL subjects had normal temporal resolution as long as the audibility was compensated, but a majority of the subjects with ANSD had poor temporal resolution in the present study.

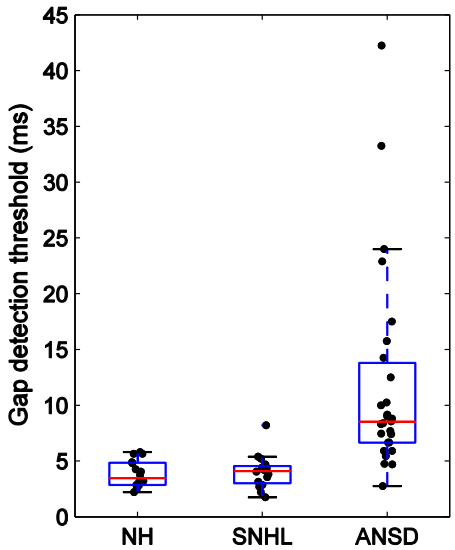


Figure 3. Temporal gap detection thresholds for subjects with NH, SNHL, and ANSD. The box represents the 25th to 75th percentiles and the horizontal line in the box represents the median. The whiskers represent the range of the data. Data points outside of the whiskers are outliers.

4 Discussion and conclusions

The fine structure (FS) cues in the speech signals play a dominant role in lexical tone perception (Xu & Pfingst, 2003; Wang et al., 2011). The temporal envelope (E) cues contribute to tone perception when the FS cues are not available (Xu et al., 2002; Kong & Zeng, 2006; Xu & Zhou, 2011). Experiment I of the present study demonstrated that the listeners with ANSD had great deficits in using the FS cues for lexical tone perception, which resulted in great difficulties in perceiving lexical tones for the ANSD listeners. Even though the ability of the listeners with ANSD to use E cues for lexical tone perception remained at a reasonable level, however, this level of performance was much lower in comparison with that of the NH listeners. The NH listeners could utilize the FS cues and achieve perfect tone perception performance. The tone perception performance in the listeners with ANSD was also lower than that in the listeners with SNHL who were still able to use the FS cues for tone perception to some extent. An earlier report has shown that the ability to use the FS cues for tone perception is negatively correlated with the degree of hearing loss in listeners with SNHL (Wang et al., 2011).

Experiments II and III evaluated the frequency resolution and temporal resolution, respectively for the subjects with SNHL and ANSD. Consistent with previous studies, the listeners with SNHL

in the present study had poor frequency resolution whereas a majority of the listeners with ANSD had normal frequency resolution. On the other hand, a majority of the listeners with ANSD had poor temporal resolution whereas the listeners with SNHL had normal temporal resolution. This may imply that poor temporal resolution rather than the frequency resolution exerts the major detrimental effects on FS cue processing for pitch perception.

5 Acknowledgements

We thank all the subjects for participating in this study and our colleagues at the Clinical Audiology Center in Beijing Tongren Hospital for helping recruit the hearing-impaired subjects. This work was funded in part by grants from the National Natural Science Foundation of China (81200754), the 2012 Beijing Nova Program (Z121107002512033), the Capital Health Research and Development of Special from the Beijing Municipal Health Bureau (2011-1017-04).

References

Apoux, F., Yoho, S.E., Youngdahl, C.L., & Healy, E.W. (2013). Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners. *Journal of the Acoustical Society of America,* 134, 2205-2212.

Glasberg, B.R., Moore, B.C.J., & Bacon, S.P. (1987). Gap detection and masking in hearing-impaired and normal-hearing subjects. *Journal of the Acoustical Society of America,* 81, 1546-1556.

Kluk, K., Moore, B.C.J. (2006). Detecting dead regions using psychophysical tuning curves: A comparison of simultaneous and forward masking. *International Journal of Audiology,* 45, 463-476.

Kong, Y.Y., & Zeng, F.G. (2006). Temporal and spectral cues in Mandarin tone recognition. *Journal of the Acoustical Society of America,* 120, 2830–2840.

Vinay, & Moore, B.C.J. (2007). Ten(HL)-test results and psychophysical tuning curves for subjects with auditory neuropathy. *International Journal of Audiology,* 46, 39-46.

Narne, V.K. (2013). Temporal processing and speech perception in noise by listeners with auditory neuropathy. *Plos One,* 8, 1-11.

Sek, A., & Moore, B.C.J. (2011). Implementation of a fast method for measuring psychophysical tuning curves. *International Journal of Audiology,* 50, 237-242.

Shannon, R.V., Zeng, F.G., & Wygonski, J. (1995). Speech recognition with primarily temporal cues. *Science,* 270, 303-304.

Smith, Z.M., Delgutte, B., & Oxenham, A.J. (2002). Chimeric sounds reveal dichotomies in auditory perception. *Nature,* 416, 87-90.

Wang, S., Mannell, R., & Xu, L. (2011). Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners. *Journal of the Association for Research in Otolaryngology,* 12, 783-794.

Xu, L., & Pfingst, B.E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception. *Journal of the Acoustical Society of America,* 114, 3024-3027.

Xu, L., Tsai, Y., & Pfingst, B.E. (2002). Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses. *Journal of the Acoustical Society of America,* 112, 247-258.

Xu, L., & Zhou, N. (2011). Tonal languages and cochlear implants. In Zeng, F.-G., Popper, A. N., & Fay, R. R. (Eds). *Auditory Prostheses: New Horizons* (pp.341-364). Springer Science+Business Media, LLC, New York.

Zeng, F.G., Oba, S., Grade, S., Sininger, Y., & Starr, A. (1999). Temporal and speech processing deficits in auditory neuropathy. *Neuroreport*, 10, 3429-3435.

Chapter 26

Interaction of Object Binding Cues in Binaural Masking Pattern Experiments

Jesko L. Verhey, Björn Lübken

Department of Experimental Audiology, Otto von Guericke University Magdeburg, Magdeburg, Germany

Steven van de Par

Acoustics group, Cluster of Excellence "Hearing4All", Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Contact author: JLV, jesko.verhey@med.ovgu.de

Abstract

Object binding cues such as binaural and across-frequency modulation cues are likely to be used by the auditory system to separate sounds from different sources in complex auditory scenes. The present study investigates the interaction of these cues in a binaural masking pattern paradigm where a sinusoidal target is masked by a narrowband noise. It was hypothesized that beating between signal and masker may contribute to signal detection when signal and masker do not spectrally overlap but that this cue could not be used in combination with interaural cues. To test this hypothesis an additional sinusoidal interferer was added to the noise masker with a lower frequency than the noise whereas the target had a higher frequency than the noise. Thresholds increase when the interferer is added. This effect is largest when the spectral interferer-masker and masker-signal distances are equal. The result supports the hypothesis that modulation cues contribute to signal detection in the classical masking paradigm and that these are analysed with modulation bandpass filters. A monaural model including an across-frequency modulation process is presented that account for this effect. Interestingly, the interferer also affects dichotic thresholds indicating that modulation cues also play a role in binaural processing.

Keywords: masking pattern, binaural masking level difference, beating, modulation processing, across-frequency process

1 Introduction

In complex acoustical environment, the auditory system makes use of various cues to group sound components from one particular sound source. These cues, sometimes referred to as object binding cues, are often associated with characteristics of the sound such as coherent level fluctuations in different frequency regions (as, e.g., observed in speech), or the location of the sources in space. Differences in location result, among others things, in interaural differences. Psychoacousticians developed experiments to study how the human auditory system makes use of the information. For example, the ability to use interaural disparities is shown in experiments on binaural masking level differences (BMLD). The BMLD is a well investigated phenomenon that signifies the reduction in masked thresholds when a tonal signal and masker have disparate interaural cues. For a diotic narrowband masker, a tonal target can be detected at a 25 dB lower

level when the interaural target phase is changed from 0 to π (e.g. van de Par and Kohlrausch, 1999). This large BMLD can only be observed when masker and target are spectrally overlapping. Several studies have shown that the BMLD reduces to rather small values for off-frequency targets (e.g., van de Par et al., 2012).

This lack of a BMLD could be interpreted as a lack of effective binaural processing in off-frequency conditions. Alternatively, it could result from monaural processing which is much better in off-frequency as compared to on-frequency conditions. The narrow-band noise masker has an envelope spectrum that predominantly has components in the range from 0 Hz until the masker bandwidth. Thus, the presence of an on-frequency tonal signal does not substantially change this envelope spectrum. When, however, the tonal signal is presented off-frequency, new, higher-frequency spectral modulation components are introduced due to the beating between tone and masker that are only masked to a small degree by the inherent masker fluctuations. This modulation cue provides for an additional lowering of off-frequency monaural detection thresholds in excess of the peripheral filtering effect. The modulation frequency selectivity of the auditory system shown in previous studies (e.g., Dau et al., 1997) indicates that the inherent modulations can be dissociated from those due to beating. The contribution of modulation cues in off-frequency masking has been demonstrated by Nitschmann & Verhey (2012) using a modulation filterbank model.

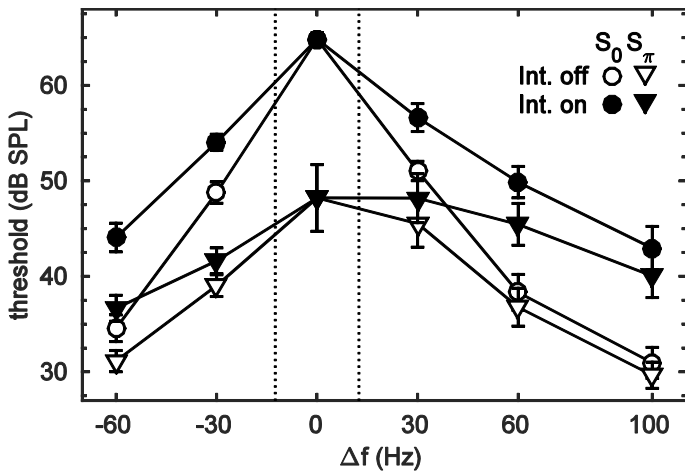


Figure 1. Masked thresholds as a function of the spectral distance between signal and masker centre frequency. Open symbols indicate data for the classical masking pattern experiment with a narrowband diotic noise masker and a sinusoidal signal. Filled symbols indicate data of a modified masking pattern experiment with an additional tonal interferer. Circles denote thresholds for the diotic condition, triangles those of the dichotic condition. The data were taken from our previous study on the effect of an interferer on masking patterns (van de Par et al., 2012).

In a previous study (van de Par et al., 2012), this monaural explanation was tested by attempting to reduce the effectiveness of the modulation cues by creating a Modulation Detection Interference (MDI) effect. To this end a tonal interferer was provided at the opposite spectral side of the masker compared to where the target was, but with the same spectral distance between target and masker as between masker and interferer. This extra interferer created beating frequencies that were equal to those that should be created by the presence of the target tone. Although these beating frequencies are then present in different auditory frequency channels, the across-frequency nature of MDI (Yost et al., 1989) implies that they should still interfere with the modulation cues used for detecting the target tone. The level of the interfering tone was chosen to be 6 dB lower than the masker level to ensure that no energetic masking should occur. Figure 1 shows the data of van de Par et al. (2012) as masking patterns with (filled symbols) and without (open symbols) interferer for the diotic (circles) and the dichotic

(triangles) conditions. As hypothesized the presence of the interferer increased the diotic thresholds substantially and had less effect on the dichotic thresholds.

In the present contribution, we investigated two aspects of the interference effect described above. First it is known that MDI exhibits some degree of frequency selectivity (Yost et al., 1989). By altering the frequency offset between the interferer tone and the narrow-band masker, the interference on the fixed sinusoidal signal should change, giving the largest interference effect when the frequency differences between the signal and interferer are the same relative to the masker centre frequency. Secondly, we will investigate to what extent the interference effect can be modelled.

2 Methods

2.1 Psychoacoustical experiment

Three masking conditions were considered. The masker of the first masking condition was a 25-Hz wide bandpass-filtered Gaussian noise with an overall level of 65 dB SPL generated using a brick-wall filter in the frequency domain. The arithmetic centre frequency f_c was 700 Hz. Apart from this noise-only condition, two other conditions were used, where an additional pure tone (interferer) was presented at 670 Hz (i.e., 30 Hz below f_c) or 600 Hz (i.e., 100 Hz below f_c). All masker and interferer components were presented diotically. The target was a sinusoid that was either presented diotically (S_0) or interaurally out of phase (S_π). The target frequency was either 730 Hz (30 Hz above f_c) or 800 Hz (100 Hz above f_c). The duration of masker, interferer and target was 500 ms including 50-ms raised cosine-ramps at onset and offset.

Masked thresholds were measured using an adaptive three-interval three-alternative forced-choice procedure. In each of the three 500-ms stimulus intervals of a trial the masker was presented. Stimulus intervals were separated by 500 ms of silence. In one randomly chosen interval of a trial the target signal was added. The subject had to indicate which interval contained the. Feedback about correctness of the answer was provided. A two-down one-up adaptive staircase procedure was used. The initial step size of 6 dB was reduced to 3 dB after the first upper reversal and to 1 dB after the second upper reversal. With this minimum step size, the run continued for another eight reversals of which the mean was used as threshold estimate. Four threshold estimates were obtained and averaged leading to the final threshold estimate for this signal and masking condition. The conditions were run in random order.

The stimuli were generated digitally with a sampling rate of 44100 Hz, converted to analog signals using a 32 bit D/A converter (RME Fireface UC), amplified (Tucker Davis KB7) and presented via headphones (Sennheiser HD650) within a sound insulated booth. Eight experienced subjects (5 male and 3 female) aged from 21 to 30 years participated. All subjects had pure tone thresholds of 10 dB HL or lower for the standard audiogram frequency.

2.2 Simulations

For the simulations a model was developed based on the model used by Nitschmann and Verhey (2012) to predict their masking pattern. The monaural front end was (i) a first-order band-pass filter with the cut-off frequencies of 0.5 and 5.3 kHz to simulate the filtering of the outer and middle ear, (ii) a bank with fourth-order gammatone filters (centre frequencies 313 – 1055 Hz) with a bandwidth of one equivalent rectangular bandwidth (ERB) and a spectral distance of adjacent filters of one ERB, (iii) white noise with a level of 7 dB SPL added to the output of each filter to simulate thresholds in quiet, (iv) half-wave rectification and a first order low-pass filter with a cut-off frequency of 1 kHz and (v) five consecutive adaptation loops to model adaptation and compression effects in the auditory system. In the monaural pathway the output of the adaptation loops was analysed by a modulation filterbank, where the highest centre frequency of the modulation filters was restricted to a quarter of the centre frequency of the auditory filter.

To simulate across-channel (AC) processes, two different versions of the monaural part of the model were used.

The first version is based on the approach of Piechowiak et al. (2007). In this approach a weighted sum of modulation filters at the output of off-frequency auditory filters was subtracted from the corresponding modulation filters at the output of the on-frequency filter. In contrast to Piechowiak et al. (2007), a multi-auditory-channel version of this approach was used. Since the weighting is essentially based on the energy in the auditory filters it will be referred to as the energy-based AC model in the following.

The second version of across-channel processing is based on the assumption that modulation in off-frequency auditory filters channels hampers the processing of the modulation filters tuned to the corresponding modulation frequency in the on-frequency auditory channel. This detrimental effect is realized by assuming that modulation frequencies adjacent to the modulation created by the masker-interferer distance could not be processed. This was implemented in an extreme form by setting the output of the modulation filters tuned to frequencies adjacent to the interferer-masker distance induced modulation frequency to zero. This approach is referred to as the modulation-based AC model.

The binaural pathway of the model is the same in both models. It is realized as an equalization-cancellation process by a low-pass filtering with a first-order low-pass filter with a cut-off frequency of 8 Hz. The final stage is an optimal detector, which calculates the cross correlation between a temporal representation of the signal at a supra-threshold level (template) with the respective actual activity pattern (Dau et al., 1997). For the simulations the same adaptive procedure was used as for the psychoacosutical experiment with the model as an artificial observer. Simulated thresholds are calculated as the mean of the estimates from 24 runs.

3 Results

Figure 2 shows average thresholds and standard errors of the eights subjects participating in the psychoacoustical experiment. Each panel shows thresholds of one signal as indicated in the top of the panel. Thresholds are shown for the three interferer conditions: “no interferer”, “interferer 30 Hz below f_c” and “interferer 100 Hz below f_c”. For the diotic 730-Hz target (top left panel), thresholds increase by 4.5 dB when an interferer was added at a distance of 30 Hz. For the spectral distance interferer-masker of 100 Hz, the threshold was 1.5 dB lower than for the spectral distance interferer-masker of 30 Hz. For the diotic 800-Hz target (top right panel), adding an interferer had a higher impact on threshold than for the lower signal frequency, as shown in Figure 1. Threshold was highest when the spectral distance masker-signal and interferer-masker threshold was the same. It was 4.5 dB lower when the interferer was positioned 30 Hz instead of 100 Hz below the masker. A similar trend was observed for the dichotic 800-Hz signal (bottom right panel) although the difference between the highest threshold for an interferer at 100 Hz below the masker and the lowest threshold without interferer was only 8 dB whereas it was 11.5 dB for the diotic 800-Hz signal. For the dichotic 730-Hz signal (bottom left panel) thresholds were about the same for the two conditions with interferer. The threshold without interferer was 1.5 dB lower than thresholds with interferer.

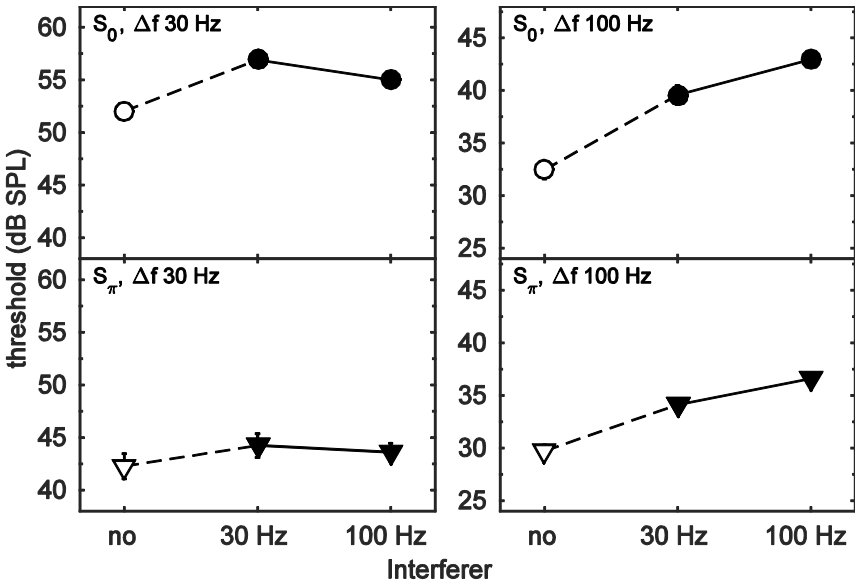


Figure 2. Measured thresholds for three different interferer conditions: without interferer (no) and with interferers at 30 Hz or 100 Hz below the masker. Each panel shows the average thresholds and standard errors of the eight subjects for one signal. The signal frequency was either 730 Hz (masker-signal Δf = 30 Hz, left panels) or 800 Hz (Δf = 100 Hz, right panels). The signal was either presented diotically (top panels) or dichotically (bottom panels).

Figure 3 shows model predictions. Since the binaural pathway is the same in both models, predictions of the two models do not differ for the dichotic signals (bottom panels). Both models predict no effect of the interferer on threshold, in contrast to the data. For the diotic 730-Hz signal (top left panel), the energy-based AC model predicted no change in threshold when the interferer was added. In contrast, the modulation-based AC model predicted an increase of about 5 dB when the interferer was added 30 Hz below the masker compared to the condition without interferer. For an interferer 100 Hz threshold was about 2 dB lower than for the interferer 30 Hz below the masker. Thus, only the modulation-based AC model predicts the effect of the interferer on thresholds for the diotic 730-Hz signal. For diotic 800-Hz signal, both models predict highest thresholds for the interferer 100 Hz below the masker and lowest for the condition without interferer. The energy-based AC model predicts an effect of 4.5 dB, i.e., considerably smaller than the 11.5 dB observed in the data. The modulation-based model predicts an effect of 14 dB. For the interferer 30 Hz below the masker, the predicted threshold was similar to the thresholds without interferer, whereas the difference between these two thresholds was 7 dB in the measured data.

4 Discussion

The behavioural results of the present study support the hypothesis that modulation cues affect detection of diotic target in masking pattern experiments and that an additional interfering tone with the same distance below (or above) the masker as the spectral distance of the signal spectrally above (or below) the masker makes the beating cue between signal and masker less usable for detection. The present study shows that the same spectral distance masker-signal and interferer-masker increases thresholds more than an interferer at a different spectral distance. This result points towards a modulation spectral selectivity as previously observed in MDI experiments.

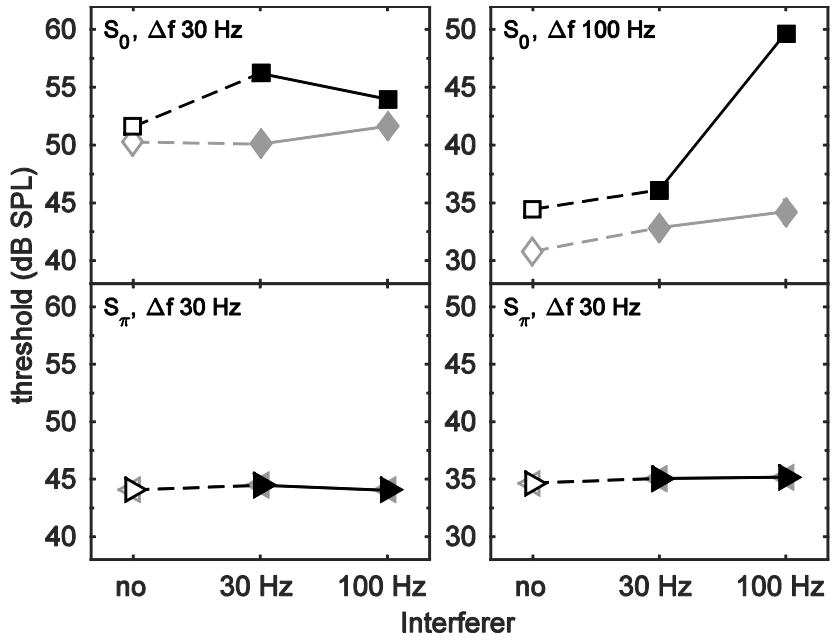


Figure 3. Same as Figure 2 but now showing simulated data. Black squares and rightward-pointing triangles indicate predictions of the modulation-based AC model. Grey diamonds and leftward-pointing triangles show the predictions of the energy-based AC model.

The present study tested an extended (energy-based AC) model of the modulation filterbank model originally designed to predict CMR by Piechowiak et al. (2007). Although this model predicted the trend for the 800-Hz signal the effect was too small. For the 730-Hz signal, no effect was predicted, in contrast to the data. Thus this energy-based AC model was unable to predict the effect of the interferer on the diotic thresholds.

The second model assumed that modulation cues with frequencies adjacent to the beating frequency between interferer and masker could not be used for signal detection. This was realized in the modulation-based AC model by setting the output of the corresponding modulation filters to zero. This approach predicted general trends of the diotic data of the present study. Among others it predicted a larger effect of the interferer for the 800-Hz signal than for the 730-Hz signal, in agreement with the data. However, there was a tendency of an overestimation of the effect for the same spectral distance masker-signal and interferer-masker and an underestimation of the effect of the interferer with a spectral distance not equal to that of signal and masker, at least for the 800-Hz signal. This indicates that the present approach was too simplistic, i.e., modulation cues may only be reduced and not abolished by the presence of the interferer, i.e., the cues may to a certain extend still be available to the subjects. In addition, not only modulation frequencies close to the beating frequency may be affected by the presence of the interferer but also other modulation frequencies. This is also consistent with MDI results where an MDI is also observed for modulation frequencies of the interferer that differ from the signal modulation frequency (Yost et al., 1989). A better match between data and simulations could be achieved by assuming weights to the output of the modulation filters that gradually increase as the distance between beating frequency (masker – interferer) increases.

As already shown in van de Par (2012), dichotic thresholds are also affected by the presence of an interferer. The present data for the dichotic 800-Hz signal indicate a similar effect of interferer-masker distance as for the diotic signal, although less pronounced. Thus, also dichotic thresholds may be influenced by a frequency-specific modulation processing. This was already hypothesized in Nitschmann & Verhey (2012). Binaural modulation frequency selectivity seems to be less accurate than monaural modulation frequency selectivity (Thompson and Dau, 2008). Thus, if binaural modulation frequency selectivity affects the dichotic off-frequency masked thresholds the mismatch of beating frequency between masker-signal and interferer-masker

should have less effect on threshold than for the diotic condition. The present data for the 800-Hz signals support this hypothesis.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, sfb trr 31).

References

Dau T., Kollmeier B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J. Acoust. Soc. Am.*, 102(5), 2892-2905.

Nitschmann, M., & Verhey, J. L. (2012). Modulation cues influence binaural masking-level difference in masking-pattern experiments. *J. Acoust. Soc. Am. Express Letters* 131(3), EL223-EL228.

Piechowiak, T., Ewert S. D., & Dau, T. (2007). Modeling comodulation masking release using an equalization-cancellation mechanism, *J. Acoust. Soc. Am.*, 121(4), 2111-2126.

Thompson, E. R., & Dau, T. (2008). Binaural processing of modulated interaural level differences. *J. Acoust. Soc. Am.*, 123(2), 1017–1029.

van de Par, S., & Kohlrausch, A. (1999). Dependence of binaural masking level differences on center frequency, masker bandwidth and interaural parameters. *J. Acoust. Soc. Am.*, Vol. 106(4), 1940-1947.

van de Par, S., Luebken, B., Verhey, J.L., & Kohlrausch A. (2012). Off-frequency BMLD: the role of monaural processing. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R.P. Carlyon, & H. E. Gockel (Eds.), *Basic Aspects of Hearing: Physiology and Perception* (pp. 293-301). Springer, New York

Yost, W. A., Sheft, S., & Opie, J. (1989). Modulation interference in detection and discrimination of amplitude modulation. *J. Acoust. Soc. Am.*, 86, 2138-2147.

Chapter 27

Speech Intelligibility for Target and Masker with a Different Spectrum

Thibaud Leclère, David Théry, Mathieu Lavandier
Laboratoire Génie Civil et Bâtiment, ENTPE, Université de Lyon, France

John F. Culling
School of psychology, Cardiff University, United Kingdom

Contact author: TL, thibaud.leclere@entpe.fr

Abstract

The speech intelligibility index (SII) calculation is based on the assumption that the effective range of signal-to-noise ratio (SNR) regarding speech intelligibility is [-15 dB; +15 dB] with an importance equally distributed over this interval. In a specific frequency band, speech intelligibility would no longer be influenced if the SNR is increased above +15 dB or decreased below -15 dB. These assumptions were tested in four experiments measuring speech reception thresholds (SRTs) with a speech target and speech-spectrum noise, while attenuating target or noise above or below 1400 Hz, with different levels of attenuation allowing the testing of different SNRs in the two bands. The results confirmed that SRT varied linearly with attenuation at low-attenuation levels and an asymptote was reached for high-attenuation levels. However, this asymptote was reached (intelligibility was not influenced by further attenuation) for different attenuation levels across conditions. The -15-dB SII limit was confirmed for high-pass filtered targets, whereas for low-pass filtered targets, intelligibility was further impaired by decreasing the SNR below -15 dB (until about -30 dB). For the high-pass and low-pass filtered noises, speech intelligibility kept improving when increasing the SNR beyond the +15 dB (up to about 40 dB). Before reaching the asymptote, a 10-dB increase of SNR obtained by filtering the noise had a larger impact on intelligibility than a corresponding 10-dB decrease of SNR obtained by filtering the target. These results question the use of the SNR range and the importance function adopted by the SII when considering sharply filtered signals.

Keywords: speech intelligibility index, speech in noise, speech intelligibility measurement

1 Introduction

Speech intelligibility in noise is strongly influenced by the relative level of the target compared to that of the noise, referred to as signal-to-noise ratio (SNR). It is well known that high SNRs will lead to better speech recognition than low SNRs. But, although SNRs can take infinite values, it is not the case for speech intelligibility. Some SNR floor and ceiling must exist, such that intelligibility would not be influenced by varying the SNR above or below these limits.

The speech intelligibility index (SII) calculation (ANSI S3.5, 1997) adopts SNR values of -15 dB and +15 dB as floor and ceiling limits respectively. This range has its origins in the work of Beranek (1947) who reported that the total dynamic range of speech is about 30 dB (in any

frequency band) by interpreting the short-term speech spectrum measurements reported by Dunn and White (1940). The aim of this study was to test these floor and ceiling SNRs by conducting four speech intelligibility experiments.

Speech Reception Thresholds (SRTs, level of the target compared to that of the noise for 50% intelligibility) were measured in the presence of a speech target and a speech-shaped noise. Target or noise was attenuated above or below 1400 Hz at different levels of attenuation in order to vary the SNR in the low or high frequency regions. SRT is expected to be increased along with the attenuation level in the case of a filtered target. Conversely, when the noise is filtered, SRT should decrease while the attenuation level is increased. In both cases, SRTs are expected to remain unchanged and form an asymptote beyond a certain attenuation level: variations of SNR would not influence speech intelligibility any longer. The knee-points (attenuation from which SRTs are no longer influenced) obtained in each experiment will be compared to those adopted by the SII standard [-15 dB; +15 dB].

General methods of the experiments are presented first, detailing the conditions and stimuli tested in this study, then the results of each experiment are presented. These results are then discussed in the last section.

2 Methods

2.1 Stimuli

Target sentences

The speech material used for the target sentences was designed by Raake & Katz (2006) and consisted of 24 lists of 12 anechoic recordings of the same male voice digitized and down-sampled here at 44.1 kHz with 16-bit quantization. These recordings were semantically unpredictable sentences in French and contained four key words (nouns, adjectives, and verbs). For instance, one sentence was “la LOI BRILLE par la CHANCE CREUSE” (the LAW SHINES by the HOLLOW CHANCE).

Maskers

Maskers were 3.8 s excerpts (to make sure that every masker is longer than the longest target sentence) of a long stationary speech-shaped noise obtained by concatenating several lists of sentences, taking the Fourier transform of the resulting signal, randomizing its phase, and finally taking its inverse Fourier transform.

Filters

Digital finite impulse response filters of 512 coefficients were designed using the host-windowing technique (Abed & Cain, 1984). High-pass (HP) and low-pass (LP) filters were used on the target or the masker at different attenuations (0 to -65 dB) depending on the experiment. The cut-off frequency was set to 1400 Hz for both HP and LP filters to achieve equal contribution from the pass and stop bands according to the SII band importance function (ANSI, 1997; Dubno et al., 2005).

2.2 Procedure

Four experiments were conducted to test each filter type (HP or LP) on each source (target or masker). Except for experiment 1 (HP target), each experiment was composed of two sub-experiments of eight conditions (because no asymptote was reached with the first set of eight conditions). Experiment 1 tested only eight conditions. In each sub-experiment, each SRT was measured using a list of twelve target sentences and an adaptive method (Brand & Kollmeier,

2002). The twelve sentences were presented one after another against a different noise of the same type. Listeners were instructed to type the words they heard on a computer keyboard after each presentation. The correct transcript was then displayed on a monitor with the key words highlighted in capital letters. Listeners identified and self-reported the number of correct key words they perceived. For the first sentence of the list, listeners had the possibility to replay the stimuli, producing an increase in the TMR of 3 dB, which was initially very low (-25 dB). Listeners were asked to attempt a transcript as soon as they believed that they could hear half of the words in the sentence. No replay was possible for the following sentences, for which the TMR was varied across sentences by modifying the target level while the masker level was kept constant at 80 dB SPL. For a given sentence, the TMR was increased if the score obtained at the previous sentence was greater than 2, it was decreased if the score was less than 2 and it remained unchanged if the score was 2. The sound level of the k^{th} ($2 < k < 12$) sentence of the list (L_k, expressed in dB SPL) was determined by Eq. 1 (Brand & Kollmeier, 2002):

$$L_k = L_{k-1} - 10 * 1.41^{-i} * (SCORE_{k-1} - 0.5) \quad (1)$$

where $SCORE_{k-1}$ is the proportion of correct key words between 0 and 1 for the sentence k-1 and i is the number of times $SCORE_{k-1}$-0.5 changed sign since the beginning of the sentence list. The SRT was taken as the mean TMR across the last eight sentences. In each sub-experiment, the SRT was measured for eight conditions presented in a pseudorandom order, which was rotated for successive listeners to counterbalance the effects of condition order and sentence lists, which were presented in a fixed sequence. Each target sentence was thus presented only once to every listener in the same order and, across a group of eight listeners, a complete rotation of conditions was achieved. In each experiment, listeners began the session with two practice runs, to get used to the task, followed by eight runs with break after four runs.

2.3 Equipment

Signals were presented to listeners over Sennheiser HD 650 headphones in a double walled soundproof booth after having been digitally mixed, D/A converted, and amplified using a Lynx TWO sound card. A graphical interface was displayed on a computer screen outside the booth window. A keyboard and a computer mouse were inside the booth to interact with the interface and gather the transcripts.

2.4 Listeners

Listeners self-reported normal hearing and French as their native language and were paid for their participation. Eight listeners took part in each sub-experiment. No listener participated at both sub-experiments within the same experiment since the target sentences used in each sub-experiment were the same.

3 Results

3.1 HP target

Figure 1 presents the SRTs measured in the presence of a HP-filtered target as a function of the filter attenuation. SRTs first increased linearly from 0 to 15-dB attenuation and then remained constant for further attenuations. Intelligibility was not disrupted any longer after filtering out the target by 15-dB attenuation. The slope of the linear increase of SRT is about 0.5 dB SRT/dB attenuation. A one-factor repeated measures analysis of variance (ANOVA) was performed on the experimental data. There was a main effect of the filter attenuation [$F(7,7) = 10.58$; $p<0.001$]. Tukey pairwise comparisons were performed on the data. Without detailing all the comparisons, none of the SRTs from 10-dB attenuation to 35-dB were significantly different, indicating a knee-point around 15-dB attenuation.

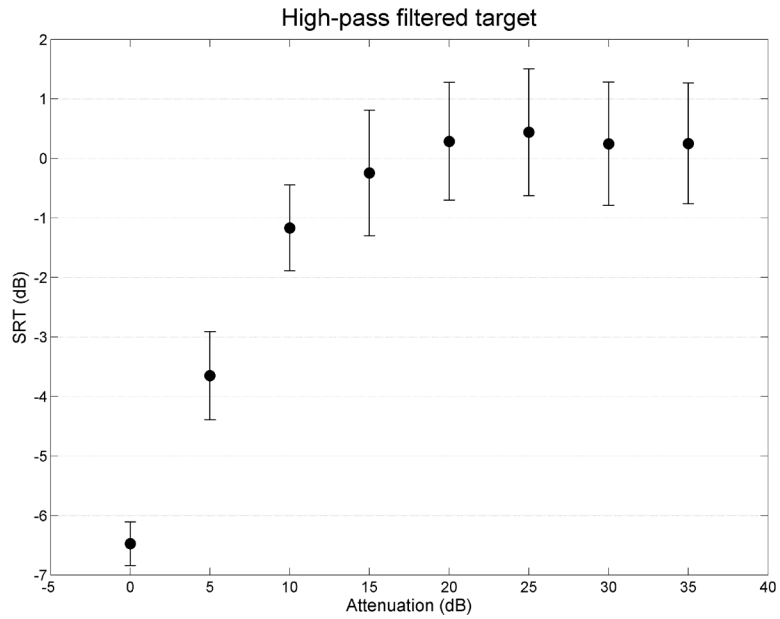


Figure 1. SRT measurements for a high-pass filtered target as a function of the filter attenuation.

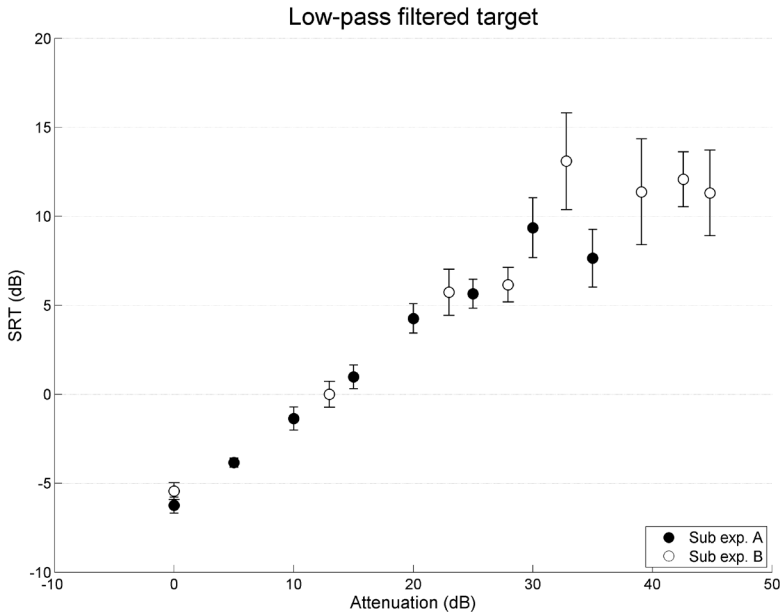


Figure 2. SRT measurements for a low-pass filtered target as a function of the filter attenuation. Filled circles correspond to the first sub-experiment while open circles correspond to the second.

3.2 LP target

SRTs measurements in the presence of a LP-filtered target are plotted in Fig. 2 as a function of the filter attenuation. As in the HP case, SRTs increased linearly with a slope of 0.5 dB SRT/dB attenuation until 30-dB attenuation. Intelligibility remained at a SRT of about 10 dB for further attenuations. A one-factor repeated measures ANOVA was performed on each sub-experiment independently. In both sub-experiments, there was a main effect of attenuation [$F(7,7) > 16.2$; $p<0.01$]. Post-hoc Tukey pairwise comparisons were performed on the data of each sub-

experiment. In sub-experiment A (filled circles), the four SRTs at the highest level of attenuation (20, 25, 30 and 35 dB) are not significantly different from each other. In sub-experiment B (open circles), the six SRTs at the highest attenuation level (23, 28, 33, 39, 42 and 45 dB) are not significantly different from each other. These analyses suggest a knee-point around 30-dB attenuation.

3.3 HP masker

SRTs measured with a HP filtered masker are presented in Fig. 3 as a function of the filter attenuation. SRTs decreased linearly with a slope of about -0.7 dB SRT/dB attenuation indicating an improvement of speech intelligibility by filtering out the low frequencies in the masker signal. Around 45-dB attenuation, SRTs stopped decreasing and presented an asymptote at about -35 dB. A one-factor repeated measures ANOVA indicated a significant main effect of the filter attenuation on speech intelligibility in both sub-experiments [$F(7,7) > 41.8$; $p < 0.01$]. Tukey pairwise comparisons were performed on the dataset of sub-experiments A and B. In sub-experiment A (filled circles), only the pairs of SRTs at the attenuations 30/25, 10/15 and 0/5 were not significantly different. All the other pairs of SRTs were significantly different from each other. In sub-experiment B (open circles), SRTs for 0 and 20-dB attenuation were significantly different from each other and all SRTs obtained for 40-dB attenuation at 40 dB and above were not significantly different from each other, suggesting a knee-point around 40-dB attenuation.

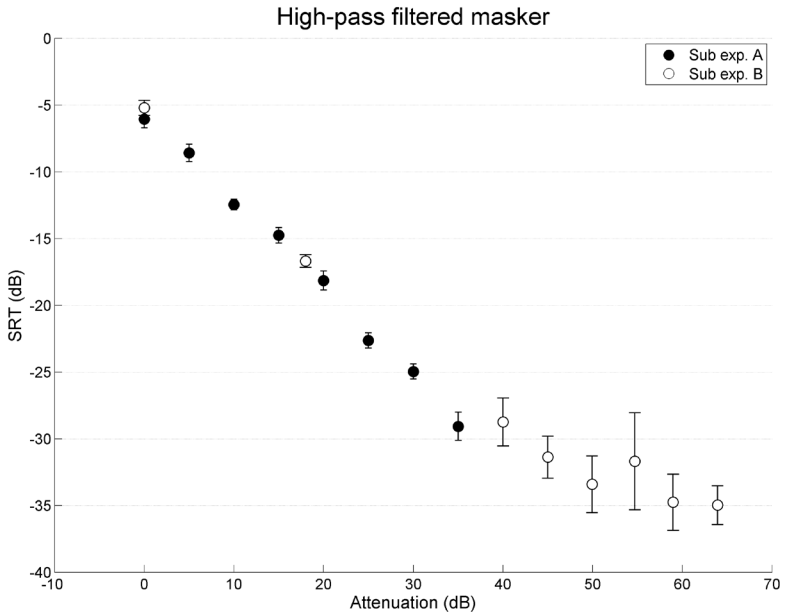


Figure 3. SRT measurements for high-pass filtered masker as a function of the filter attenuation. Filled circles correspond to the first sub-experiment while open circles correspond to the second.

3.4 LP masker

Figure 4 presents the SRTs measurements obtained in the presence of a LP filtered masker as a function of the filter attenuation. As in the HP case, SRTs linearly decreased with attenuation until 40-dB attenuation. For further attenuations, SRTs were constant at about -35 dB. The slope of the linear decrease of SRTs was about -0.7 dB SRT/dB attenuation. A one-factor repeated measures ANOVA was performed on the data from each sub-experiment. There was a main effect of the filter attenuation [$F(7,7) > 44.5$; $p < 0.01$], which was further interrogated by performing Tukey pairwise comparisons on the dataset from each sub-experiment. In sub-experiment A (filled circles), all pairs of SRTs were different from each other except for those

corresponding to an attenuation at 38 dB and above. In sub-experiment B (open circles), none of the SRTs between 34-dB and 65-dB attenuation were significantly different. SRTs obtained for lower attenuations were all significantly different from each other. These pairwise comparisons suggested a knee-point located around an attenuation of 40 dB.

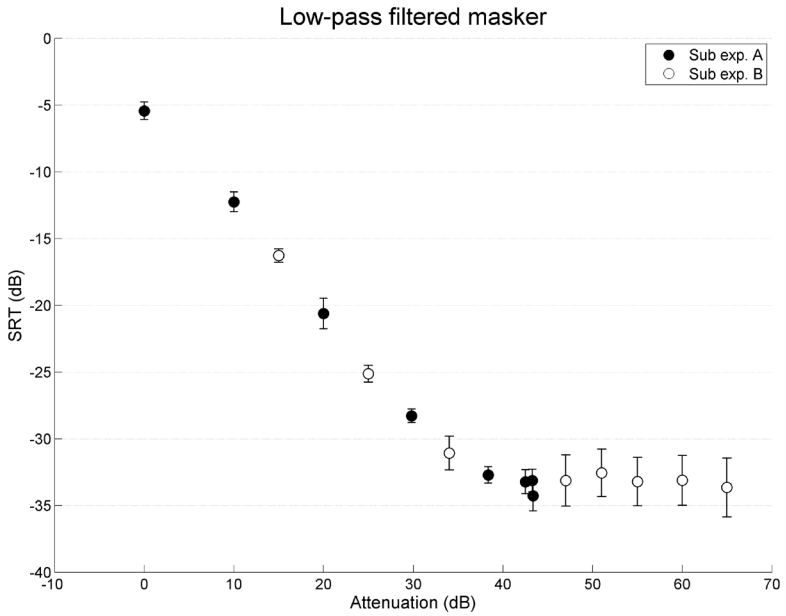


Figure 4. SRT measurements of low-pass filtered masker as a function of the filter attenuation. Filled circles correspond to the first sub-experiment while open circles correspond to the second.

4 Discussion

In each experiment, speech intelligibility was strongly influenced by the SNR in a specific band. However, SRTs were not affected any longer beyond a certain attenuation level (knee-point) and remained constant. The knee-point values observed in these experiments are not in agreement with the SII assumptions, except when the target was HP-filtered. In this specific case, the results suggested a knee-point around 15-dB attenuation. The results from the other experiments suggested a larger range of SNRs contributing to speech intelligibility [-25 dB; 40 dB]. When the target is filtered, the knee-point seemed to be frequency dependent in contrast to the SII approach which proposes the same SNR range in each frequency band. It is worth noting that the SII standard was not designed for sharply filtered speech or sharply filtered noise such as the stimuli used in the present study. The SII is then not expected to predict the data presented here but other values for the ceiling and floor SNRs than those proposed by the standard are needed to describe speech intelligibility of sharply filtered signals.

Changing the filter type (HP or LP) had no influence on the slope of the linear increase (or decrease) of SRTs before reaching the asymptote, confirming that the chosen low and high frequency regions contributed equally to speech intelligibility. But, the results showed that filtering the masker seemed to have a larger impact on speech intelligibility than filtering the target. SRTs were not symmetrical as the SNR was varied up or down with the same amount. The benefit was greater when the filtered part of the noise decreased rather than when the filtered part of the target increased. This result questions the uniformly distributed importance over the [-15; +15] interval adopted by the SII and suggests that a greater importance should be attributed to positive SNRs compared to negative SNRs.

Studebaker and Sherbecoe (2002) derived intensity importance functions from intelligibility scores in five frequency bands. The functions they obtained differed from the one used in the SII calculation. Their functions are frequency-dependent and nonlinear along SNRs, which is in agreement with the findings of the present study. However, the derivation of their results yielded importance functions defined between SNRs of -15 dB and +45 dB. The results observed here suggested a floor SNR of -15 dB only when the target was HP-filtered. For the other conditions, lower SNRs need to be considered. The importance function of Studebaker and Sherbecoe also attributes negative contribution to speech intelligibility in the presence of very high SNR (> 30 dB on average across frequency bands). These negative contribution might be due to high sound levels used in their study which could have led to poorer word recognition scores (Dubno et al., 2005). Their derived importance functions take into account both the effect of SNR and absolute speech level. It would be preferable to separate the influence of these two parameters. Their approach could however be inspiring in order to propose simple frequency dependent SNR-importance functions, allowing to model the SRTs measured in the present study.

Acknowledgments

The authors would like to thank all listeners who took part in the experiments. This work was performed within the Labex CeLyA (ANR-10-LABX-0060/ANR-11-IDEX-0007).

References

Abed, A.-E. H. M. & Cain, G. D. (1984). The Host Windowing Technique for FIR Digital Filter Design. IEEE Transactions on Acoustics Speech and Signal Processing, *ASSP.32*, 683-694.

ANSI S3.5 (**1997**). Methods for calculation of the speech intelligibility index. American National Standards Institute, New York .

Beranek, L. L. (1947). The design of speech communication systems. Proceedings of the Institute of Radio Engineers, 35, 880-890

Brand, T. & Kollmeier, B. (2002). Efficient adaptive procedures for thresholds and concurrent slope estimates for psychophysics and speech intelligibility tests. Journal of the Acoustical Society of America, 111(6), 2801-2810.

Dubno, J. R.; Horwitz, A. R. & Ahlstrom, J. B. (2005). Recognition of filtered words in noise at higher-than-normal levels: Decreases in scores with and without increases in masking. Journal of Acoustical Society of America, 118(2), 923-933.

Dunn, H. K. and White, S. D. (1940). Statistical measurements on conversational speech. Journal of the Acoustical Society of America, 11, 278-288

Raake, A. & Katz, B. F. G. (2006). SUS-based method for speech reception threshold measurement in French. Proceedings of Languages Resources and Evaluation Conference, 2028-2033.

Studebaker, G. A. and Sherbecoe, R. L. (2002). Intensity-importance functions for bandlimited monosyllabic words. Journal of the Acoustical Society of America, 111, 1422-1436

Chapter 28

Dynamics of Cochlear Nonlinearity

Nigel P. Cooper, Marcel van der Heijden
Erasmus MC, Rotterdam, The Netherlands

Contact author: M. van der Heijden - m.vanderheyden@erasmusmc.nl

Abstract

Dynamic aspects of cochlear mechanical compression were studied by recording basilar membrane (BM) vibrations evoked by tone pairs ("beat stimuli") in the 11-19 kHz region of the gerbil cochlea. The frequencies of the stimulus components were varied to produce a range of "beat rates" at or near the characteristic frequency (CF) of the BM site under study, and the amplitudes of the components were balanced to produce near perfect periodic cancellations, visible as sharp notches in the envelope of the BM response. We found a compressive relation between instantaneous stimulus intensity and BM response magnitude that was strongest at low beat rates (e.g., 10 - 100 Hz). At higher beat rates, the amount of compression reduced progressively (i.e. the responses became linearized), and the rising and falling flanks of the response envelope showed increasing amounts of hysteresis; the rising flank becoming steeper than the falling flank. This hysteresis indicates that cochlear mechanical compression is not instantaneous, and is suggestive of a gain control mechanism having finite attack and release times. In gain control terms, the linearization that occurs at higher beat rates occurs because the instantaneous gain becomes smoothened, or low-pass filtered, with respect to the magnitude fluctuations in the stimulus. In terms of peripheral processing, the linearization corresponds to an enhanced coding, or decompression, of rapid amplitude modulations. These findings are relevant both to those who wish to understand the underlying mechanisms and those who need a realistic model of nonlinear processing by the auditory periphery.

Keywords: basilar membrane, hair cell, compression, gain control

1 Introduction

Basilar-membrane (BM) vibrations in healthy cochleae grow compressively with the intensity of stationary sounds such as tones and broad band noise (de Boer & Nuttall, 1997; Rhode, 1971; Robles & Ruggero, 2001). Similarly compressive nonlinearities also affect mechanical responses to pulsatile stimuli such as clicks (e.g. Robles et al., 1976), and in two-tone experiments (Robles et al., 1997), where they reputedly act on a cycle-by-cycle basis to generate distinctive patterns of intermodulation distortion products. The extent to which a compressive input-output relation holds when the intensity of a stimulus fluctuates in time has not been studied systematically, however, and the possibility that the cochlea operates more like an automatic gain control (AGC) system than an instantaneously nonlinear system remains feasible (van der Heijden, 2005). In the present study, we investigate the dynamics of the cochlea's compressive mechanism(s) by observing basilar membrane (BM) responses to balanced, beating pairs of low-level tones near the recording site's characteristic frequency (CF).

2 Methods

Sound-evoked BM vibrations were recorded from the basal turn of the cochlea in terminally anesthetised gerbils, using techniques similar to those reported previously (Versteegh & van der Heijden, 2012). Acoustic stimuli included: (i) inharmonic multi-tone complexes (zwuis stimuli), which were used to characterise the tuning properties of each site on the BM (including the site's CF); and (ii) "beating" pairs of inharmonic near-CF tones, with component levels adjusted to produce near-perfect periodic-cancellations in the response envelopes at individual sites on the BM (cf. Figure 1).

3 Results

3.1 Time-Domain Observations

BM responses to well balanced, "beating" two-tone stimuli had heavily compressed envelopes, as illustrated in Figure 1. The amount of compression was quantified by comparing the shapes of the observed response envelopes with those that would be predicted in a completely linear system (illustrated by the dashed blue and red lines in Figs. 1B and C, respectively). When scaled to have similar peak magnitudes, the observed envelopes exceeded the linear predictions across most of the beat cycle (cf. Figs. 1B, C). Expressing the ratios of the observed and predicted envelopes as "normalised gains" (Fig. 1D), the envelope's maximum gain always occurred near (but not exactly at) the instant of maximal cancellation between the two beating tones. This is as expected in a compressive system, where (by definition) "gain" decreases with increasing intensity, or increases with decreasing intensity.

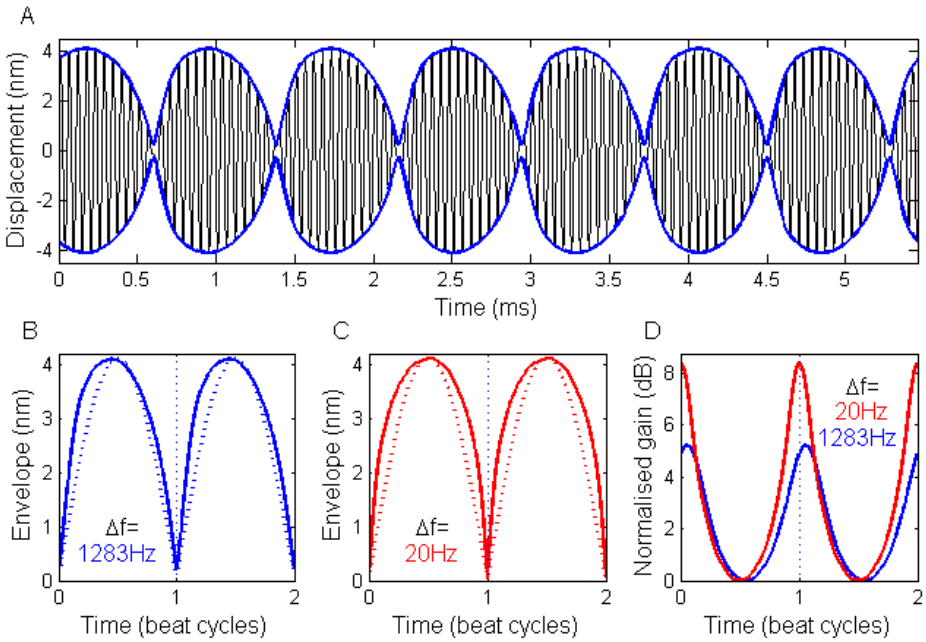


Figure 1. (A) BM responses to near-CF beat stimuli. f_1=17373 Hz, 30dB SPL, f_2=18656 Hz, 33dB SPL, $\Delta f=f_2-f_1$=1283 Hz. (B) Observed (solid blue line, from A) and linearly predicted (dashed blue line) response envelopes for Δf=1283 Hz. (C) Observed (solid red line) and predicted (dashed red line) response envelopes for Δf=20 Hz. (D) Normalised gain (=observed/linearly-predicted envelope) for Δf=20 Hz (red) and 1283 Hz (blue). Experiment RG15641, CF=18 kHz.

3.2 Temporal Asymmetry

At low beat rates (e.g. for $\Delta f=f_2-f_1$ ~10-160 Hz), observed response envelopes appeared symmetric in the time domain (cf. Fig. 1C). At higher rates, however, the rising flanks of the response envelopes became steeper than the falling flanks (cf. Figs. 1A, B). This temporal asymmetry reflects hysteresis, and is reminiscent of the output of a gain control system with separate attack and decay time-constants.

3.3 "Gain Control" Dynamics

Three further response characteristics (in addition to the hysteresis) became more pronounced at higher beat rates: (1) the point at which the observed envelopes exhibited their maximal gains (relative to the linearly predicted responses) occurred later and later in the beat cycle, (2) the overall amount of envelope gain decreased, and (3) the peakedness of this gain within the beat-cycle decreased (compare the red and blue curves in Figs. 1D). These characteristics were confirmed and quantified by spectral analysis of the envelopes' temporal gain curves, as illustrated in Figure 2.

The left panels of Figure 2 show spectral decompositions of the envelope gain functions (cf. Fig. 1D) for a wide range of stimulus beat rates. Response energy at each harmonic of the beat rate decreases as the beat rate increases (Fig. 2A), suggesting that envelope gain is subjected to a low-pass filtering (or smoothing) process before it is observed on the BM: this is the spectral counterpart to the decreased gain at higher beat rates (point 2) referred to above. Response energy also decreases with increasing harmonic number in Figure 2 (i.e. the coloured curves in Fig. 2A are more closely spaced at low beat rates, and wider apart at higher rates): this is the spectral counterpart of the decreased peakedness of the temporal gain curves (point 3) referred to above.

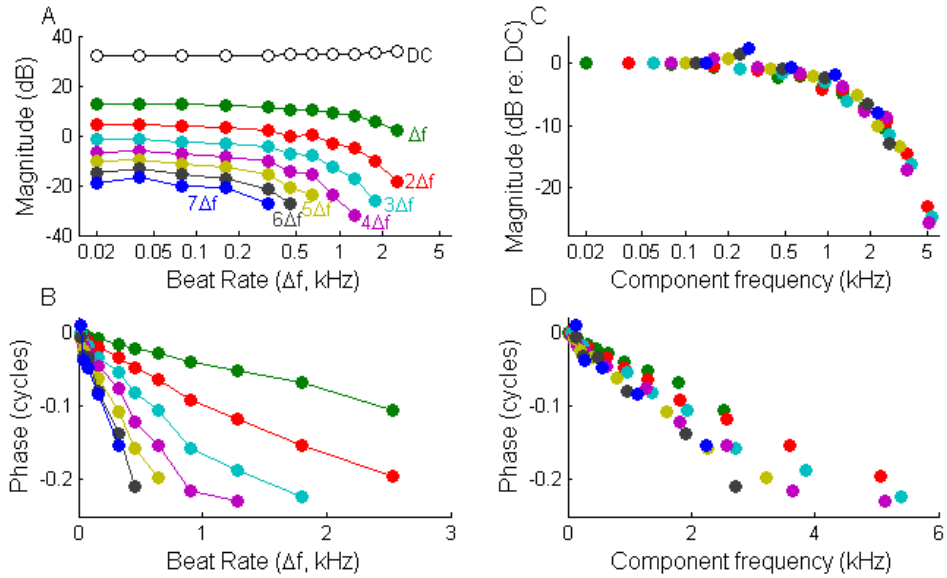


Figure 2. Spectral analysis of beat response envelope gain in experiment RG15641. (A) Magnitude and (B) phase of envelope gain functions (cf. Fig. 1D) as a function of beat rate ($\Delta f=f_2-f_1$). Coloured curves represent the contributions of the different harmonics of Δf (labelled in A) to the envelope gain functions at each beat rate. (C) Magnitude and (D) phase of the envelope transfer function derived from A and B. Coloured symbols from A and B are plotted at the actual frequencies of the spectral components (i.e. at f=n.Δf) for the whole family of beat rates tested (20-2560 Hz), and magnitudes are normalized re: the quasi-static (Δf=20Hz) harmonics.

The low-pass filtering apparent in the spectral analysis of the beat responses can be characterised most conveniently by deriving a transfer function for the response envelope gain, as illustrated in Figures 2C and D. This shows the amount of energy in the gain function at each harmonic of the beat rate, after normalization by the BM's quasi-static (i.e. DC) input-output function. In the case of Figure 2C, we have used the 20 Hz beat rate data to normalize the data at higher frequencies, and plotted each of the harmonics at its own absolute frequency (i.e. at f=n.Δf, where n is the harmonic number). This reveals a relatively simple low-pass transfer function, with a -3 dB cut-off at ~1 kHz.

The energy at each harmonic of the beat rate also occurs later in the beat cycle as the beat rate increases, as shown in Figures 2B and D (negative phases in Fig. 2 indicate phase-lags). The phases in Fig. 2B scale almost proportionally with harmonic number, suggesting that (on average) the observed envelope gain curves lag the effective stimuli (the linearly predicted envelopes) by a constant amount of time. This delay is approximately 50 μs for most of our data (i.e. equating to 0.2 cycles of phase-lag accumulating in a 4 kHz range, as shown in Fig. 2D). These observations are the spectral counterparts of the delays observed in the instants of maximum envelope gain mentioned in section 3.1 (cf. Fig. 1D, where the red curve peaks 50 μs after the null in the observed response envelope).

3.4 Linearization of Function

The observed decreases in envelope gain and in gain peakedness with increasing beat rate act synergistically to reduce the amount of compression exhibited in the BM response envelopes at high beat rates. This "linearization" is illustrated directly in Figure 3 by plotting the response envelopes as a function of the instantaneous intensity of the beating stimulus. The decreases in gain are observed as clockwise rotations in the linear coordinates of Fig. 3A, and as downward shifts on the logarithmic plots of Fig. 3B. The linearization that accompanies the decreases in gain is seen as an increase in the threshold above which compression is seen (e.g. the blue 20 Hz data of Fig. 3B start to deviate from linearity above ~10 dB SPL and ~0.3 nm, whereas the steepest, rising flank of the black 1280 Hz data remains nearly linear up to ~25 dB SPL and 2 nm). One consequence of this linearization is that BM's response dynamics can actually enhance the encoding of dynamic stimuli, within certain limits. However, this "enhancement" comes at the price of reduced overall sensitivity.

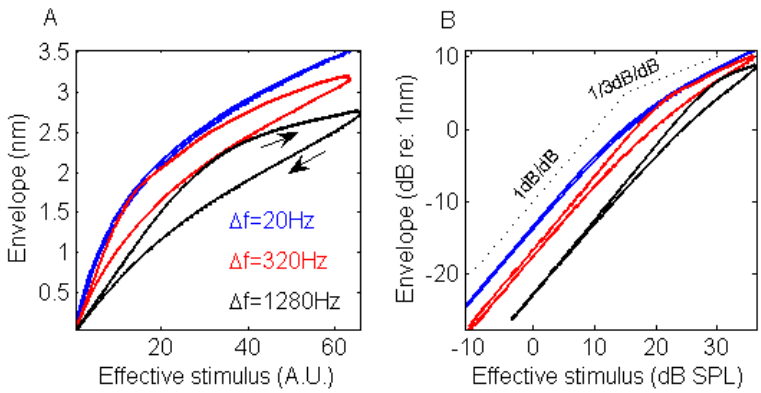


Figure 3. Response linearization with increasing beat rate (experiment RG14612, CF=15 kHz). Instantaneous envelope input-output functions for three beat stimuli on (A) linear and (B) logarithmic scales. Arrows in A distinguish rising and falling phases of 1280Hz data. Dashed lines in B show linear (1 dB/dB) and compressive (0.333 dB/dB) growth rates.

4 Discussion

The mechanical nonlinearity of the mammalian cochlea is clearly not instantaneous. The results of the current investigation show that dynamic changes in stimulus intensity can only be followed accurately up to a rate of ~1 kHz in the basal turn of the gerbil cochlea. This limitation appears to be produced by a low-pass gain control filter of some kind: this could be a mechanical filter, imparted by the load on a force generating element (such as a motile outer hair cell, or OHC), or an electrical filter (such as the OHC's basolateral membrane, cf. Housley & Ashmore, 1992). The filter could even involve a combination of multiple, coupled mechanisms: preliminary analysis suggests that the gain at one site on the BM may be controlled remotely, presumably by elements (e.g., OHCs) that are distributed more basally along the BM.

Our observation of hysteresis in the BM's response to a dynamic stimulus is not entirely new – similarly asymmetric envelope responses have been reported in studies using amplitude modulated tonal stimuli (e.g., Rhode & Recio, 2001). Our observation of a distinct delay to the appearance of compression (or gain) in the response envelopes is new, however. We currently believe that it is impossible to produce such a delay without resorting to a non-instantaneous form of gain control, such as an AGC. Other (non-AGC) types of system which produce level-dependent group delays can easily mimic the envelope asymmetry (hysteresis) effects that we observe, but (to the best of our knowledge) they cannot produce an actual delay in the "effect" of the nonlinearty. If this delay really does signify the presence of an AGC in the cochlea, it may well prove significant for the coding of a wide range of stimuli, and have consequences for numerous psychophysical phenomena.

Acknowledgments

Supported the Netherlands Organization for Scientific Research, ALW 823.02.018.

References

De Boer, E. & Nuttall, A. L. (1997). The mechanical waveform of the basilar membrane. I. Frequency modulations ("glides") in impulse responses and cross-correlation functions. *The Journal of the Acoustical Society of America*, *101*(6), 3583–3592. http://doi.org/10.1121/1.418319

Housley, G. D. & Ashmore, J. F. (1992). Ionic currents of outer hair cells isolated from the guinea-pig cochlea. *The Journal of Physiology*, *448*, 73–98.

Rhode, W. S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. *The Journal of the Acoustical Society of America*, *49*(4), Suppl 2:1218+.

Rhode, W. S. & Recio, A. (2001). Multicomponent stimulus interactions observed in basilar-membrane vibration in the basal region of the chinchilla cochlea. *The Journal of the Acoustical Society of America*, *110*(6), 3140–3154.

Robles, L., Rhode, W. S. & Geisler, C. D. (1976). Transient response of the basilar membrane measured in squirrel monkeys using the Mössbauer effect. *The Journal of the Acoustical Society of America*, *59*(4), 926–939.

Robles, L. & Ruggero, M. A. (2001). Mechanics of the Mammalian Cochlea. *Physiological Reviews*, *81*(3), 1305–1352.

Robles, L., Ruggero, M. A. & Rich, N. C. (1997). Two-tone distortion on the basilar membrane of the chinchilla cochlea. *Journal of Neurophysiology*, *77*(5), 2385–2399.

Van der Heijden, M. (2005). Cochlear gain control. *The Journal of the Acoustical Society of America*, *117*(3 Pt 1), 1223–1233.

Versteegh, C. P. C. & van der Heijden, M. (2012). Basilar Membrane Responses to Tones and Tone Complexes: Nonlinear Effects of Stimulus Intensity. *JARO: Journal of the Association for Research in Otolaryngology*, *13*(6), 785–798. http://doi.org/10.1007/s10162-012-0345-0

Chapter 29

Responses of the Human Inner Ear to Low-Frequency Sound

Markus Drexl, Eike Krause, Robert Gürkov

German Center for Vertigo and Balance Disorders (IFB), Grosshadern Medical Centre, University of Munich, Munich, Germany
and
Department of Otorhinolaryngology, Head and Neck Surgery, Grosshadern Medical Centre, University of Munich, Munich, Germany

Lutz Wiegrebe

Department Biology II, University of Munich, Martinsried, Germany

Contact author: M.D., markus.drexl@med.uni-muenchen.de

Abstract

The perceptual insensitivity to low frequency (LF) sound in humans has led to an underestimation of the physiological impact of LF exposure on the inner ear. It is known, however, that intense, LF sound causes cyclic changes of indicators of inner ear function after LF stimulus offset, for which the term "Bounce" phenomenon has been coined. Here, we show that the mechanical amplification of hair cells (OHCs) is significantly affected after the presentation of LF sound. First, we show the Bounce phenomenon in slow level changes of quadratic, but not cubic, distortion product otoacoustic emissions (DPOAEs). Second, Bouncing in response to LF sound is seen in slow, oscillating frequency and correlated level changes of spontaneous otoacoustic emissions (SOAEs). Surprisingly, LF sound can induce new SOAEs which can persist for tens of seconds. Further, we show that the Bounce persists under free-field conditions, i.e. without an in-ear probe occluding the auditory meatus. Finally, we show that the Bounce is affected by contralateral acoustic stimulation synchronized to the ipsilateral LF sound. These findings clearly demonstrate that the origin of the Bounce lies in the modulation of cochlear amplifier gain. We conclude that activity changes of OHCs are the source of the Bounce, most likely caused by a temporary disturbance of OHC calcium homeostasis. In the light of these findings, the effects of long-duration, anthropogenic LF sound on the human inner ear require further research.

Keywords: Cochlea, calcium homeostasis, cochlear amplifier, outer hair cells, inner hair cells, Bounce phenomenon, otoacoustic emissions

1 Introduction

For decades, low-frequency sound, i.e. sound with frequencies lower than 250 Hz (Berglund et al., 1996), has been considered to largely bypass the inner ear even at intense levels, simply because human hearing thresholds for frequencies below 250 Hz are relatively high. Recent evidence from animal models shows that physiological cochlear responses to LF sound are even

larger than those evoked by equal-level, higher frequencies in the more sensitive range of hearing (Salt et al., 2013). No data for human subjects are available, but, considering the higher sensitivity of humans for LF sounds, similar results can be expected (Salt et al., 2013).

Hirsh and Ward (1952) observed temporary deteriorations of human absolute thresholds about 2 minutes after presenting subjects with an intense, non-traumatic LF tone. Later on, the term 'Bounce' was used to describe bimodal changes in absolute thresholds starting with a sensitisation period followed by an about equal-duration temporary desensitization (Hughes, 1954).

Perceptual thresholds essentially reflect the sensitivity of inner hair cells (IHCs), which are functionally coupled to inner ear fluids (Nowotny & Gummer, 2006; Guinan, 2012). IHCs are therefore sensitive to basilar-membrane velocity, which decreases with decreasing frequency. OHCs, in contrast, are mechanically linked to both the basilar membrane and the tectorial membrane. OHCs are therefore sensitive to basilar membrane displacement (Dallos et al., 1982; Dallos, 1986), which does not decrease with decreasing stimulus frequency. Thus, OHCs are more sensitive to LF sound than IHCs and it is this difference in LF sensitivity, which contributes to the LF limit of sound perception (Salt & Hullar, 2010; Salt et al., 2013). In humans, non-invasive recordings of DPOAEs allow indirect access to OHC function while SOAEs represent, for ears that exhibit them, a more direct and very sensitive marker of OHC function. SOAEs are narrowband acoustic signals which are spontaneously emitted by the inner ear in the absence of acoustic stimulation. Human SOAEs persist over years and are relatively stable in both frequency and level (Burns, 2009).

Here, we use both DPOAE and SOAE measurements to assess LF-induced changes of cochlear physiology and active sound amplification. Specifically, we monitored the sound level and frequency of DPOAEs and SOAEs before and after the exposure to a 90 s LF sinusoid with 30 Hz and a level of 80 dBA (120 dB SPL). Both the sound level and the exposure duration were controlled to be within the exposure limits for normal working conditions as regulated by the European Commission Noise at Work Directive 2003/10/EC.

2 Methods

Data were collected from young adult normal hearing subjects. The ethics committee of the University Hospital of the Ludwig-Maximilians University Munich, Germany, in agreement with the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans, approved the procedures, and all subjects gave their written informed consent. An Etymotic Research 10C DPOAE probe system was used for recording of OAEs. The LF tone (30 Hz sine wave, 120 dB SPL, 90 s, including 0.1 s raised-cosine ramps) was supplied by a separate loudspeaker (Aurasound NSW1-205-8A). This loudspeaker was connected via a 50 cm polyethylene tube (inner diameter 1 mm) and the foam ear tip of the ER-10C DPOAE probe so that it faced the tympanic membrane. The loudspeaker was driven by a Rotel RB-960BX power amplifier. Stimulation to the contralateral ear was provided by an Etymotic Research 4PT earphone, which was sealed into the contralateral ear canal with foam ear tips. The earphone was driven by the headphone amplifier of the audio interface (RME audio Fireface UC, fs=44.1 kHz) which was programmed (using MatLab and the HörTech SoundMexPro audio toolbox) for synchronous stimulation and recording of all required inputs and outputs.

3 Results

3.1 Effect of LF sound exposure on DPOAEs

The effect of the 80 dBA LF exposure on quadratic (QDP) and cubic (CDP) OAEs is shown in Figure 1A and B, respectively. In 14 out of 20 tested subjects, the LF exposure induced a

subsequent increase of the QDP level lasting for about 60 to 90 s (see Fig. 1A). QDP levels increased with a median of 3.4 dB. In most cases, this QDP increase was followed by a similar QDP decrease (median: -2.4 dB), at about 120-150 s post-exposure. This decrease slowly recovered to pre-exposure QDP levels. The median duration of the overall oscillatory change of the QDP level was 214 s.

In many cases it was also possible to extract CDP levels from the same recording (albeit f2/f1 ratios were optimized to achieve maximum QDP levels). Typically, we observed no significant changes of CDP level after LF sound exposure (Fig. 1B).

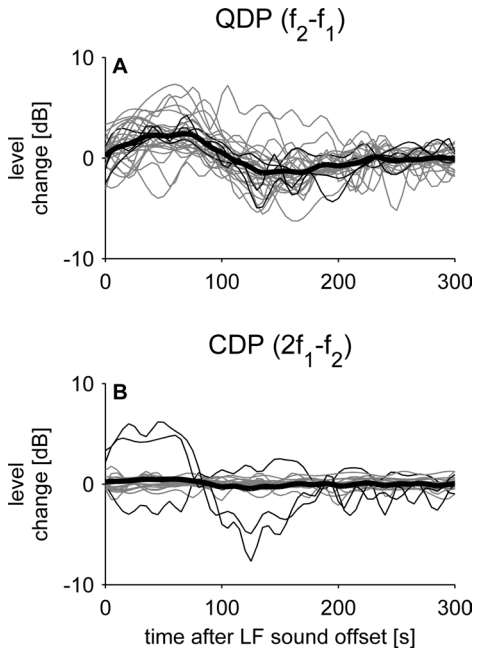


Figure 1. Effect of LF sound exposure on quadratic distortion products (QDP, A, N = 28 from 14 subjects) and cubic distortion products (CDP, B, N = 22 from 11 subjects). QDPs oscillate significantly after LF offset (A) whereas CDPs are rarely affected (B). Bold lines represent median DPOAE measures, grey lines are individual DPOAE measures. Modified after Drexl et al. (2014).

3.2 Effect of LF sound exposure on SOAEs

We recorded 80 SOAEs from 27 ears of 16 young, normal-hearing subjects. The median SOAE sound levels were 0.6 dB SPL (first and third quartiles, -4.5 dB SPL; 4.0 dB SPL) with a signal-to-noise ratio of 16.6 dB (11.6 dB, 23.5 dB).

After LF sound stimulation, 56 of these 80 SOAEs increased in both sound level and frequency. This increase was followed by a decrease of both level and frequency relative to pre-exposure (see Fig. 2A). In 10 of the 80 pre-exposure SOAEs, we observed an inverted pattern with an initial level and frequency decrease, followed by a level and frequency increase.

SOAE level- and frequency oscillations were fitted with an (inverted-phase) underdamped sinusoidal oscillation. The period of the fitted sinusoid was 257 s (202 s, 294 s) for the level time course and 252 s (215 s, 367 s) for the frequency time course. The time constant of the damped sinusoid for level changes was 120 s (76 s, 157 s) and for frequency changes 94 s (58 s, 141 s). SOAE frequency changes amounted to 5 Cent (4 Cent, 9 Cent) with peak values of 25 Cent. Relative to the SOAE frequency in the control condition, the frequency showed initial maximum increases of 4 Cent (3 Cent, 7 Cent), followed by maximum decreases of 1 Cent (0 Cent, 2 Cent).

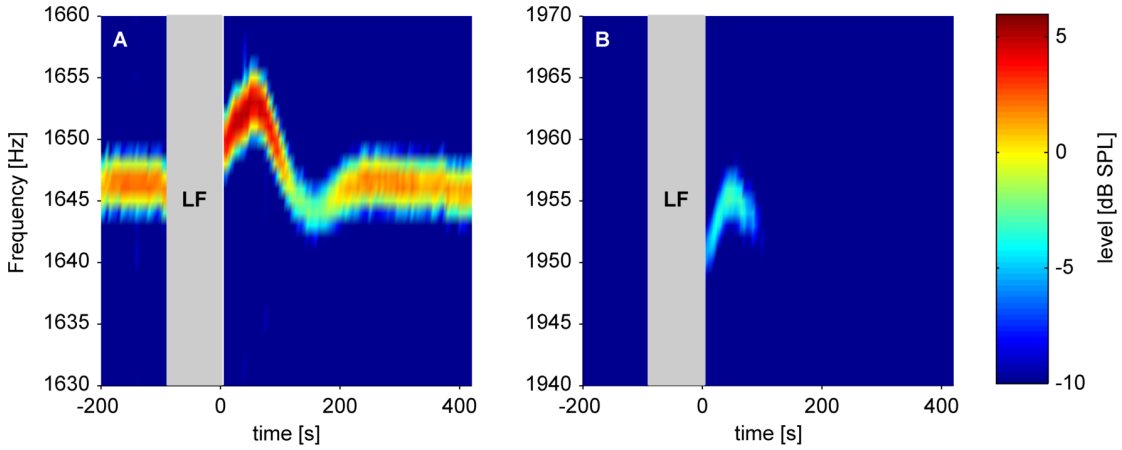


Figure 2. A: Exemplary level and frequency changes of an SOAE recorded pre- and post LF sound exposure. The grey bar indicates the presentation of the LF stimulus (30 Hz, 80 dBA, 90 s). B: Same as in a, but for a new SOAE which only appears for a short period after LF exposure. Modified after Kugler et al. (2014).

Seventeen of 21 tested subjects revealed an overall of 56 new SOAEs, which had not been measurable before LF stimulation (see Fig. 2B). These new SOAEs were characterized by an initial level and frequency increase, qualitatively similar to the pre-existing SOAEs. Comparable to the enhancing half cycle of Bouncing SOAEs, their level and frequency oscillated before they disappeared into the noise floor. The duration of the level and frequency changes was 67.5 s (47.5 s, 90 s). New SOAEs started to arise within 12.5 s (5 s, 25 s) after LF sound offset and reached a level maximum at 50 s (35 s, 62.5 s) after LF offset. The maximum SOAE level was -0.3 dB SPL (-4.1 dB SPL, 4.9 dB SPL) with a signal to noise ratio of 13.8 dB (11.9 dB, 17.6 dB). The difference between the new SOAE frequency maximum and minimum was 4 Cent (1 Cent, 6 Cent). The time course of level and frequency changes was almost identical and maximum level and frequency changes coincided.

3.3 SOAE Bouncing in the free sound field

Although the observed pattern of synchronized SOAE frequency- and amplitude changes is incompatible with the SOAE bouncing being elicited by changes in middle ear impedance, it is conceivable that bouncing may be only seen in the closed sound field where the auditory meatus is blocked by the OAE probe. Here we recorded SOAEs in the open meatus using an ER10C probe microphone fitted to the meatus via an about 8 cm silicon tube (2.8 mm outer diameter) which did not block the meatus. The tip of the tube was positioned about 10 mm in front of the tympanum. LF exposure was provided by two powerful custom-made subwoofers. Subjects lay on a deck chair in a long basement corridor at a point where standing waves in the corridor maximized the sound level at 30 Hz. LF exposure was 118 dB SPL for 90 s. Photos of LF stimulation apparatus and probe-tube placement are shown in Figure 3A, B.

Both ears of 45 young, normal-hearing subjects were screened for SOAEs. 33 subjects showed at least one SOAE in one ear. Overall we could record in the open meatus about 52% of those SOAEs detectable in a sound-attenuated room and a closed-ear recording technique. The remaining 48% were not significantly above the much higher noise floor of the free-field, open-meatus measurement.

Exemplary measurements of both permanent and transient SOAEs are shown in Figure 3C and D in the same format as Figure 2. Indeed many of those 48 % SOAEs that had been initially identified in the closed-meatus measurements, but could no longer be detected in the open meatus, appeared directly after the LF exposure for a short time period, before falling again below the noise floor (Fig. 3D). These data clearly show that Bouncing of SOAEs can indeed be elicited by free-field exposure to LF sound sources of natural or anthropogenic origin.

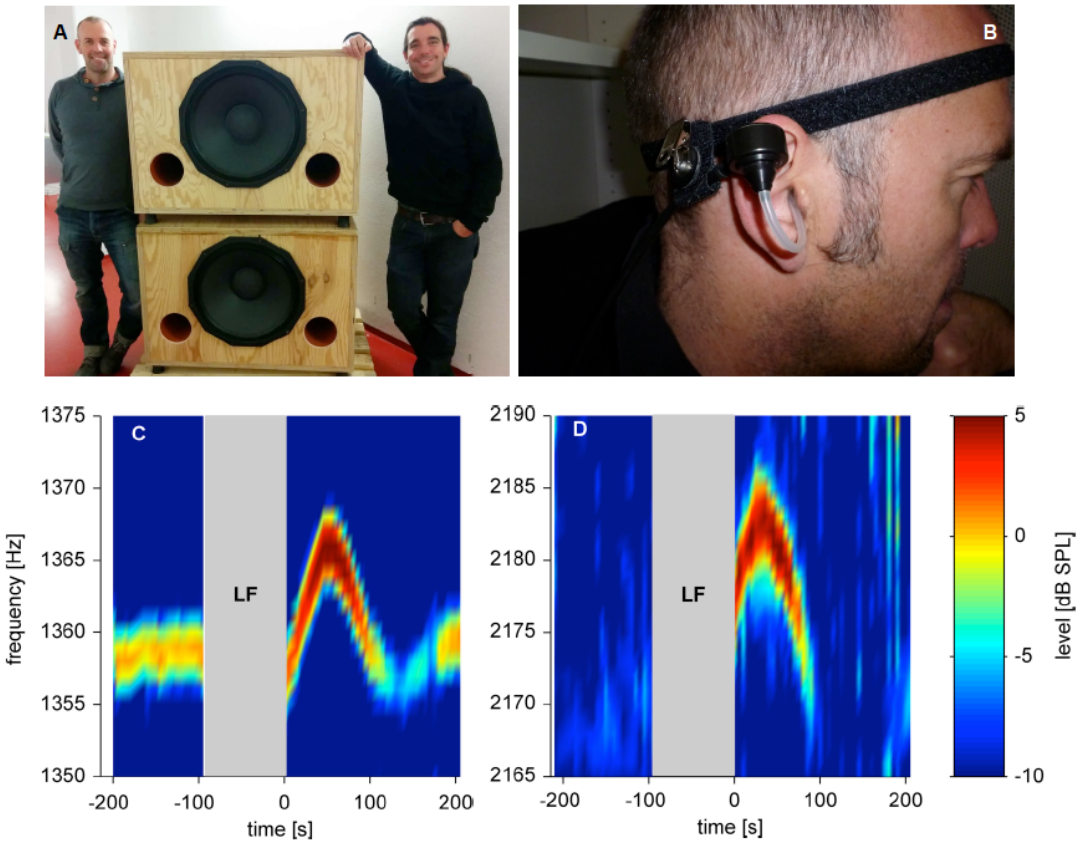


Figure 3. LF free-field stimulation subwoofers (A) and positioning of the SOAE recording probe in the open ear canal (B). Exemplary recordings in the open meatus of a pre-existing SOAE (C) and a transient SOAE (D) after free-field exposure to a 118 dB SPL, 30 Hz, 90 s LF sound (grey bar).

3.4 Effect of contralateral acoustic stimulation (CAS)

(Patuzzi, 2011) suggested that large receptor-potentials elicited by low-frequency stimulation produce a net Ca^{2+} influx.. The Bounce presumably reflects an underdamped, homeostatic readjustment of increased Ca2+ concentrations and related gain changes after low-frequency sound offset. Here, we tested this hypothesis by activating the medial olivocochlear (MOC) efferent system during presentation of the Bounce-evoking LF sound. The MOC system is known to modulate OHC Ca^{2+} concentrations (Sridhar et al., 1997) and receptor potentials (Fex, 1967) and therefore it should modulate the characteristics of the Bounce. Here, CAS was provided simultaneously to the (ipsilateral to the observed SOAE) LF exposure. The CAS consisted of a 90 s, bandpass-filtered Gaussian noise (100 Hz – 8 kHz) presented at 65 or 70 dB SPL.

CAS is well known to suppress ipsilaterally recorded SOAEs during presentation and SOAEs quickly recover after CAS offset within less than one second (Zhao & Dhar, 2010, 2011). Due to the duration of our analysis segments (5 s), the SOAEs already fully recovered from the CAS exposure within the first analysis segment. Consequently, we found no SOAEs fulfilling our criteria for the Bounce or indeed any other significant changes in the CAS control recordings. When the CAS was presented simultaneously with the ipsilateral LF tone, however, Bouncing of permanent SOAEs after LF offset changed significantly. Exemplary time courses of a preexisting and a transient SOAE are shown in Figure 4A and B, respectively. While in the reference recording (red) the preexisting SOAE showed a significant biphasic Bounce, presentation of a 65 dB SPL (blue) or 70 dB SPL (green) CAS together with the ipsilateral LF tone clearly affected the magnitude of the Bounce.

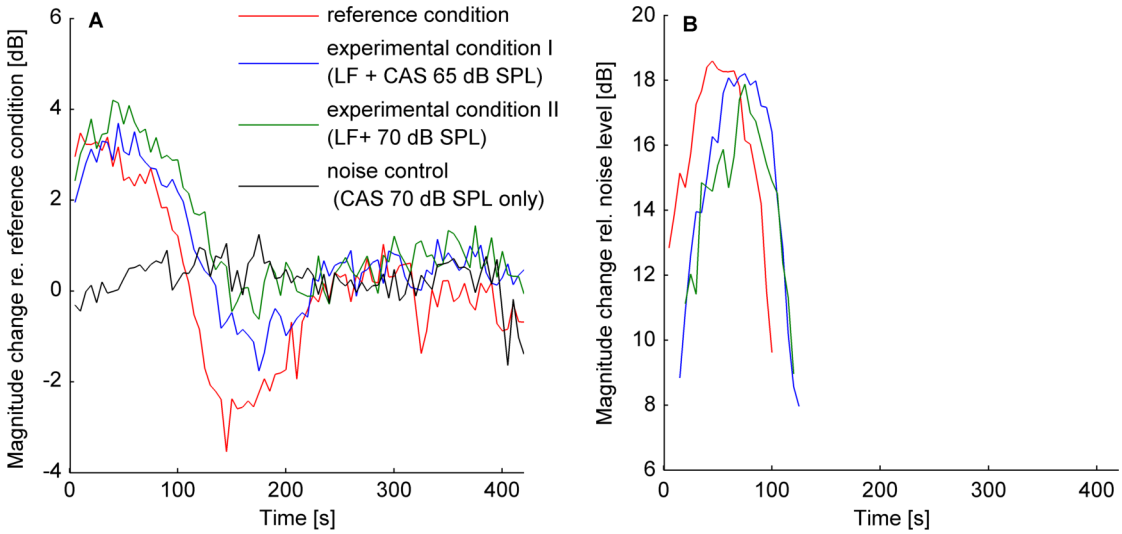


Figure 4. A preexisting SOAE (A) and a new, transient SOAE (B) after LF exposure (red) and after LF exposure witch simultaneous CAS at 65 dB SPL (blue) or 70 dB SPL (green), and after CAS exposure alone (black).

Overall, temporary level reductions of preexisting SOAEs were less pronounced with CAS than without (Wilcoxon signed rank test, p=0.085 and 0.007 for CAS levels of 65 and 70 dB, respectively), while the SOAE level increases remained fairly unchanged. This resulted in a less symmetrical shape of the SOAE Bounce compared to the reference recording. Consequently, the time constants of the fitted function with CAS shortened significantly.

4 Discussion

The current biophysical experiments reveal a significant effect of LF sound exposure on the inner ear, and specifically on the function of OHCs. The observed effects are reflected in significant changes of quadratic distortion products, but importantly, in changes of SOAEs, which (for the ears that exhibit them) allows for a least-invasive evaluation of inner ear function.

As the low-frequency stimuli used in this study are intense, mechanisms associated with acoustic overexposure and recovery thereof could in principle be responsible for the Bounce phenomenon. Temporary or permanent damage to OHCs represent an unlikely cause, as recovery from acoustic overexposure with sounds in the sensitive range of hearing is typically monotonic, does not oscillate and no hypersensitivity can be seen. It has been shown, however, that intense sound stimulation of the cochlea in the isolated temporal bone preparation increases intracellular Ca^{2+} level of hair cells and supporting cells (Fridberger et al. 1998; Jacob et al. 2013). Acoustic injury consists of a plethora of structural and metabolic changes to the cochlea, with structural damage possibly masking more subtle (and possibly oscillating) metabolic changes of cochlear sensitivity caused by the rise of intracellular Ca^{2+} levels.

The observed effects support the hypothesis that OHC Ca^{2+}homeostasis is the source of the Bounce: Patuzzi (2011) suggested that the Bounce is a direct result of OHC activation by LF sound, rather than a secondary effect caused by modulation of stria vascularis activity, as the LF sound-induced endocochlear potential modulation during the Bounce (Kirk and Patuzzi 1997) could suggest. Patuzzi (2011) hypothesized that large, LF-induced receptor potentials in OHCs are the underlying cause of cochlear sensitivity oscillations associated with the Bounce phenomenon. These LF OHC receptor potentials are not attenuated by the low-pass characteristic of the OHC membrane. Patuzzi (2011) postulated that large, LF receptor potentials activate voltage-sensitive, L-type Ca^{2+} channels in the OHC membrane. This results in an increase of intracellular Ca^{2+} in OHCs. Ca^{2+}-induced Ca^{2+}release and -uptake mechanisms, with different

time courses, can then cause slow, underdamped oscillations of OHC cytosolic Ca^{2+} concentrations, modulating the gain of the cochlear amplifier (Patuzzi 2011).

The increased damping (corresponding to decreased decay time constants) we observed in the CAS experiments indicates that processes re-adjusting the Ca^{2+} overshoot may accelerate due to the activation of the medial olivo-cochlear bundle by the CAS. Even while the LF sound is on, the LF-induced Ca^{2+} concentration changes presumably also undergo oscillations (Patuzzi 2011). In contrast, the slow efferent effect (Sridhar et al., 1997) can cause a constant Ca^{2+} release from OHC internal stores, the lateral cisternae, while CAS is on. We hypothesize that this constant influx of Ca^{2+} may help to accelerate Ca^{2+}-dependent Ca^{2+}-uptake and thus, a quicker recovery of the system.

In summary, the current experiments reveal a pronounced effect of LF exposure on the active mechanisms in the inner ear, as they are mediated by OHCs. Considering that the current LF exposure was limited to 90 s, it is unclear how a longer-duration LF exposure may affect the system.

5 Acknowledgments

The authors wish to acknowledge the contributions of Margarete Überfuhr, David Laubender and Kathrin Kugler, who carried out the experimental procedures this chapter is based on.

This work was funded by a grant (EO 0901) from the German Ministry of Science and Education to the German Center for Vertigo and Balance Disorders (IFB), project TR-F9, to the authors.

References

Berglund, B., Hassmen, P., & Job, R. F. (1996). Sources and effects of low-frequency noise. [Comparative Study Research Support, Non-U S Gov't Review]. *Journal of the Acoustical Society of America, 99*(5), 2985-3002.

Burns, E. M. (2009). Long-term stability of spontaneous otoacoustic emissions. *Journal of the Acoustical Society of America, 125*(5), 3166-3176. doi: 10.1121/1.3097768

Dallos, P. (1986). Neurobiology of cochlear inner and outer hair cells: intracellular recordings. [Research Support, U.S. Gov't, P.H.S.]. *Hearing Research, 22*, 185-198.

Dallos, P., Santos-Sacchi, J., & Flock, A. (1982). Intracellular recordings from cochlear outer hair cells. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Science, 218*(4572), 582-584.

Drexl, M., Uberfuhr, M., Weddell, T. D., Lukashkin, A. N., Wiegrebe, L., Krause, E., & Gurkov, R. (2014). Multiple indices of the 'bounce' phenomenon obtained from the same human ears. *J Assoc Res Otolaryngol, 15*(1), 57-72. doi: 10.1007/s10162-013-0424-x

Fex, J. (1967). Efferent inhibition in the cochlea related to hair-cell dc activity: study of postsynaptic activity of the crossed olivocochlear fibres in the cat. *Journal of the Acoustical Society of America, 41*(3), 666-675.

Guinan, J. J. (2012). How are inner hair cells stimulated? Evidence for multiple mechanical drives. *Hearing Research, 292*(1-2), 35-50.

Hirsh, I., & Ward, W. (1952). Recovery of the auditory threshold after strong acoustic stimulation. *Journal of the Acoustical Society of America, 24*, 131.

Hughes, J. R. (1954). Auditory Sensitization. *Journal of the Acoustical Society of America, 26*(6), 1064-1070. doi: Doi 10.1121/1.1907450

Kugler, K., Wiegrebe, L., Grothe, B., Kössl, M., Gürkov, R., Krause, E., & Drexl, M. (2014). Low-frequency sound affects active micromechanics in the human inner ear. [10.1098/rsos.140166]. *Royal Society Open Science, 1*(2).

Nowotny, M., & Gummer, A. W. (2006). Nanomechanics of the subtectorial space caused by electromechanics of cochlear outer hair cells. *Proceedings of the National Academy of Sciences of the United States of America, 103*(7), 2120-2125. doi: 10.1073/pnas.0511125103

Patuzzi, R. (2011). Ion flow in cochlear hair cells and the regulation of hearing sensitivity. *Hearing Research, 280*(1-2), 3-20. doi: DOI 10.1016/j.heares.2011.04.006

Salt, A. N., & Hullar, T. E. (2010). Responses of the ear to low frequency sounds, infrasound and wind turbines. *Hearing Research, 268*(1-2), 12-21. doi: S0378-5955(10)00312-6 [pii]

10.1016/j.heares.2010.06.007

Salt, A. N., Lichtenhan, J. T., Gill, R. M., & Hartsock, J. J. (2013). Large endolymphatic potentials from low-frequency and infrasonic tones in the guinea pig. *Journal of the Acoustical Society of America, 133*(3), 1561-1571. doi: 10.1121/1.4789005

Sridhar, T. S., Brown, M. C., & Sewell, W. F. (1997). Unique postsynaptic signaling at the hair cell efferent synapse permits calcium to evoke changes on two time scales. *J Neurosci., 17*(1), 428-437.

Zhao, W., & Dhar, S. (2010). The effect of contralateral acoustic stimulation on spontaneous otoacoustic emissions. [Research Support, N.I.H., Extramural]. *J Assoc Res Otolaryngol, 11*(1), 53-67. doi: 10.1007/s10162-009-0189-4

Zhao, W., & Dhar, S. (2011). Fast and slow effects of medial olivocochlear efferent activity in humans. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *PloS One, 6*(4), e18725. doi: 10.1371/journal.pone.0018725

Chapter 30

Suppression Measured from Chinchilla Auditory-Nerve-Fiber Responses Following Noise-Induced Hearing Loss: Adaptive-Tracking and Systems-Identification Approaches

Mark Sayles[1,3], Michael K. Walls[1], Michael G. Heinz[1,2,*]

[1] *Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, USA*
[2] *Weldon School of Biomedical Engineering, Purdue University, West Lafayette, USA*
[3]*Laboratory of Auditory Neurophysiology, Campus Gasthuisberg, O&N II, Leuven, Belgium*

Contact author: M.G.H., mheinz@purdue.edu

Abstract

The compressive nonlinearity of cochlear signal transduction, reflecting outer-hair-cell function, manifests as suppressive spectral interactions; e.g., two-tone suppression. Moreover, for broadband sounds, there are multiple interactions between frequency components. These frequency-dependent nonlinearities are important for neural coding of complex sounds, such as speech. Acoustic-trauma-induced outer-hair-cell damage is associated with loss of nonlinearity, which auditory prostheses attempt to restore with, e.g., "multi-channel dynamic compression" algorithms.

Neurophysiological data on suppression in hearing-impaired (HI) mammals are limited. We present data on firing-rate suppression measured in auditory-nerve-fiber responses in a chinchilla model of noise-induced hearing loss, and in normal-hearing (NH) controls at equal sensation level. Hearing-impaired (HI) animals had elevated single-fiber excitatory thresholds (by ~20-40 dB), broadened frequency tuning, and reduced-magnitude distortion-product otoacoustic emissions; consistent with mixed inner- and outer-hair-cell pathology. We characterized suppression using two approaches: adaptive tracking of two-tone-suppression threshold (62 NH, and 35 HI fibers), and Wiener-kernel analyses of responses to broadband noise (91 NH, and 148 HI fibers). Suppression-threshold tuning curves showed sensitive low-side suppression for NH and HI animals. High-side suppression thresholds were elevated in HI animals, to the same extent as excitatory thresholds. We factored second-order Wiener-kernels into excitatory and suppressive sub-kernels to quantify the relative strength of suppression. We found a small decrease in suppression in HI fibers, which correlated with broadened tuning. These data will help guide novel amplification strategies, particularly for complex listening situations (e.g., speech in noise), in which current hearing aids struggle to restore intelligibility.

Keywords: auditory nerve, suppression, frequency-dependent nonlinearity, masking, chinchilla, Wiener kernel, singular-value decomposition, threshold tracking, frequency tuning, hearing impairment, noise exposure, acoustic trauma, cochlear hearing loss, auditory-brainstem response, otoacoustic emissions.

1 Introduction

Frequency-dependent cochlear signal-transduction nonlinearities manifest as suppressive interactions between acoustic-stimulus components (Sachs & Kiang, 1968; de Boer & Nuttall, 2002; Versteegh & van der Heijden, 2012, 2013). The underlying mechanism is thought to be saturation of outer-hair-cell receptor currents (Geisler et al., 1990; Cooper, 1996). Despite the importance of suppression for neural coding of complex sounds (e.g., Sachs & Young, 1980), relatively little is known about suppression in listeners with cochlear hearing loss (Schmiedt et al., 1990; Miller et al., 1997; Hicks & Bacon, 1999). Here we present preliminary data on firing-rate suppression, measured with tones, and with broadband noise, from ANFs in chinchillas (*Chinchilla laniger*), following noise-induced hearing loss. Quantitative descriptions of suppression in the hearing-impaired auditory periphery will help guide novel amplification strategies, particularly for complex listening situations (e.g., speech in noise), in which current hearing aids struggle to improve speech intelligibility.

2 Methods

2.1 Animal Model

Animal procedures were under anesthesia, approved by Purdue University's IACUC, and followed NIH-issued guidance. Evoked-potential and DPOAE measures, and noise exposures, were with ketamine (40 mg/kg i.p.) and xylazine (4 mg/kg s.c.). Single-unit neurophysiology was done with sodium pentobarbital (boluses: 5-10 mg/hr. i.v.).

Noise Exposure & Hearing-Loss Characterization

Two groups of chinchillas are included: normal-hearing controls (NH), and hearing-impaired (HI) animals with a stable, permanent, sensorineural hearing loss. Prior to noise exposure, we recorded auditory brainstem responses (ABRs) and distortion-product otoacoustic emissions (DPOAEs) from the HI group, verifying "normal" baseline status.

ABRs were recorded in response to pure-tone bursts, at octave-spaced frequencies between 0.5 and 16 kHz. Threshold was determined according to statistical criteria (Henry et al., 2011). DPOAE stimulus primaries (f_1 and f_2) were presented with equal amplitude (75-dB SPL), and with a constant f_2/f_1 ratio of 1.2. f_2 varied between 0.5 and 12 kHz, in 2-semitone steps.

Animals were exposed to either a 500-Hz-centered octave-band noise at 116-dB SPL for 2 hours, or a 2-kHz-centered 50-Hz-wide noise at 114-115-dB SPL for 4 hours. They were allowed to recover for 3-4 weeks before an acute single-unit experiment. At surgery, ABR and DPOAE measures were repeated prior to any additional intervention [for NH animals, these were their only ABR and DPOAE measurements].

Single-Unit Neurophysiology

The auditory nerve was approached via a posterior-fossa craniotomy. ANFs were isolated using glass pipettes with 15-25 MΩ impedance. Spike times were recorded with 10-μs resolution. Stimulus presentation and data acquisition were controlled by MATLAB programs interfaced with hardware modules (TDT and National Instruments). For HI animals, ANF characteristic frequency (CF) was determined by the high-side-slope method of Liberman (1984).

2.2 Stimuli

Adaptive-Tracking: Suppression-Threshold Tuning

This two-tone suppression (2TS) adaptive-tracking technique is based on Delgutte (1990). Stimuli were sequences of 60-ms duration supra-threshold tones at CF, with or without a second tone at the suppressor frequency (FS; Fig. 1). For each FS frequency-level combination, 10 repetitions of the two-interval sequence were presented. The decision to increase or decrease sound level was made based on the mean of the 2nd through 10th of these two-interval comparisons. The algorithm sought the lowest sound level of FS which *reduced* the response to the CF tone by 20 spikes s^{-1}.

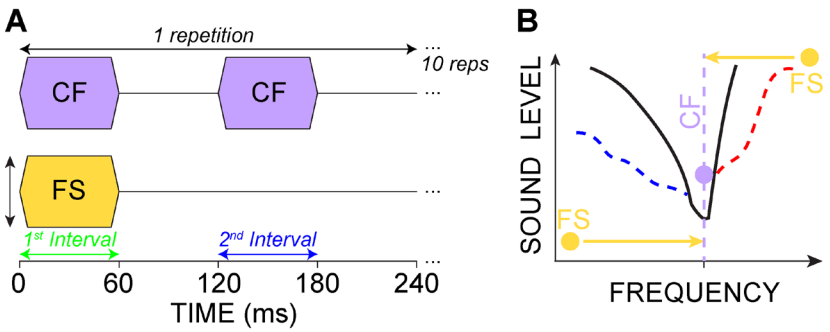


Figure 1. Schematized stimulus paradigm: adaptive tracking. (B) Black line, excitatory tuning curve; Red and blue dashed lines, high- and low-side suppression-threshold curves, respectively.

Systems-Identification Approach

Using spike-triggered characterization, we probed suppressive influences on firing rate in response to broadband-noise stimulation (e.g., Lewis & van Dijk, 2004; Schwartz et al., 2006). Our implementation is based on singular-value decomposition of the second-order Wiener kernel (h_2) in response to 16.5-kHz bandwidth, 15-dB SL noise (Lewis et al., 2002a, 2002b; Recio-Spinoso et al., 2005). We collected ~10-20K spike times per fiber.

2.3 Analyses

Second-Order Wiener Kernels

For each fiber, h_2 was computed from the second-order cross-correlation between the noise-stimulus waveform $x(t)$ and N spike times. The spike-triggered cross correlation was sampled at 50 kHz, and with maximum time lag τ of 10.2 ms (m=512 points) for CFs >3 kHz, or 20.4 ms (m=1024 points) for CFs <3 kHz. $h_2(\tau_1, \tau_2)$ is calculated as

$$\frac{N_0}{A^2}[R_2(\tau_1,\tau_2) - \phi_{xx}(\tau_2 - \tau_1)]$$

where τ_1 and τ_2 are time lags, A the instantaneous noise power, N_0 the mean firing rate in spikes s^{-1}, and $R_2(\tau_1,\tau_2)$ the second-order reverse-correlation function calculated as

$$\frac{1}{N}\sum_{i=1}^{N} x(t_i - \tau_1)x(t_i - \tau_2)$$

with t_i the i^{th} spike time and $\phi_{xx}(\tau_2 - \tau_1)$ the stimulus autocorrelation matrix. So computed, h_2 is an m-by-m matrix with units of spikes·s^{-1}·Pa^{-2} (Recio-Spinoso et al., 2005). We used singular-

value decomposition to parse h_2 into excitatory ($h_{2\varepsilon}$) and suppressive ($h_{2\sigma}$) sub-kernels (Lewis et al., 2002a, 2002b; Lewis & van Dijk, 2004; Rust et al., 2005; Sneary & Lewis, 2007).

Excitatory and Suppressive Sub-Kernels

Using the MATLAB function *svd*, h_2s were decomposed as

$$h_2 = \mathrm{USV}$$

where U, S and V are m-by-m matrices. The columns of U and rows of V are the *left* and *right singular vectors*, respectively. S is a diagonal matrix, the nonzero values of which are the weights of the corresponding-rank vectors. The decomposition can be rephrased as

$$h_2 = \sum_{j=1}^{m} k_j u_j v_j$$

where u_j and v_j are column-vector elements of U and V, respectively, and k_j is the signed weight calculated as

$$k_j = sgn\big[u_j(j)v_j(j)\big]s_j$$

where sgn is the signum function, $u_j(j)$ is the j^{th} element of the j^{th} left singular vector, $v_j(j)$ is the j^{th} element of the j^{th} right singular vector, and s_j is the j^{th} element of the nonzero diagonal of S.

Positively and negatively weighted vectors are interpreted as excitatory and suppressive influences, respectively (Rust et al., 2005; Schwartz et al., 2006; Sneary & Lewis, 2007). However, mechanical suppression, adaptation, and refractory effects may all contribute to what we term "suppressive" influences.

To determine statistical significance of each weighted vector, we re-computed h_2 from 20 different spike-train randomizations, conserving the first-order inter-spike-interval distribution. Based on this bootstrap distribution, weights were expressed as z-scores, and any vector with $|z| > 3$ and rank ≤20 was considered significant. We calculated a normalized excitatory-suppressive ratio as

$$R_{\varepsilon,\sigma} = \frac{\sum_{j=1}^{N_\varepsilon} |z_{j\varepsilon}| - \sum_{j=1}^{N_\sigma} |z_{j\sigma}|}{\sum_{j=1}^{N_\varepsilon} |z_{j\varepsilon}| + \sum_{j=1}^{N_\sigma} |z_{j\sigma}|}$$

for N_ε significant excitatory vectors and N_σ significant suppressive vectors. This normalized ratio varies from 1 (only excitation, no suppression), through 0 (equal excitation and suppression), to -1 (only suppression and no excitation: in practice $R_{\varepsilon,\sigma}<0$ does not occur).

Spectro-Temporal Receptive Fields

Based on Lewis & van Dijk (2004), and Sneary & Lewis (2007), we estimated the spectro-temporal receptive field (STRF) from h_2. These STRFs indicate the timing of spectral components of the broadband-noise stimulus driving either increases or decreases in spike rate. Moreover, the STRF calculated from the whole h_2 kernel is the sum of excitatory and suppressive influences. Therefore, we also determined STRFs separately from $h_{2\varepsilon}$ and $h_{2\sigma}$ to assess the tuning of excitation and suppression (Sneary & Lewis, 2007).

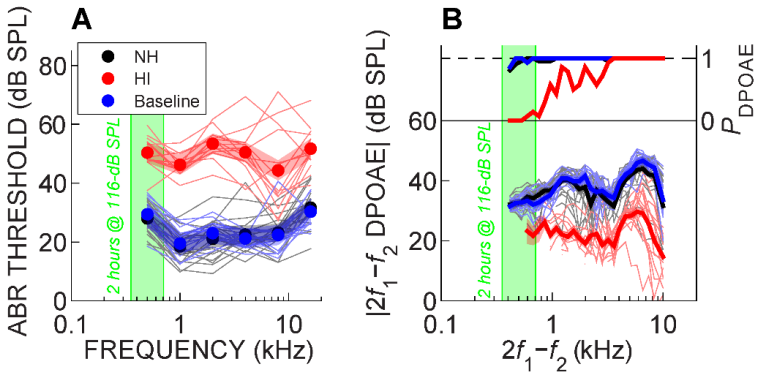


Figure 2. Audiometric characterization. (A), ABR thresholds. Thin lines, individual-animal data; symbols, within-group least-squares-mean values; shading, S.E.M; green area, noise-exposure band. (B), *Lower plot*, DPOAE magnitude; *Upper plot*, probability of observing a DPOAE above the noise floor; thick lines, within-group least-squares means; shading, S.E.M.

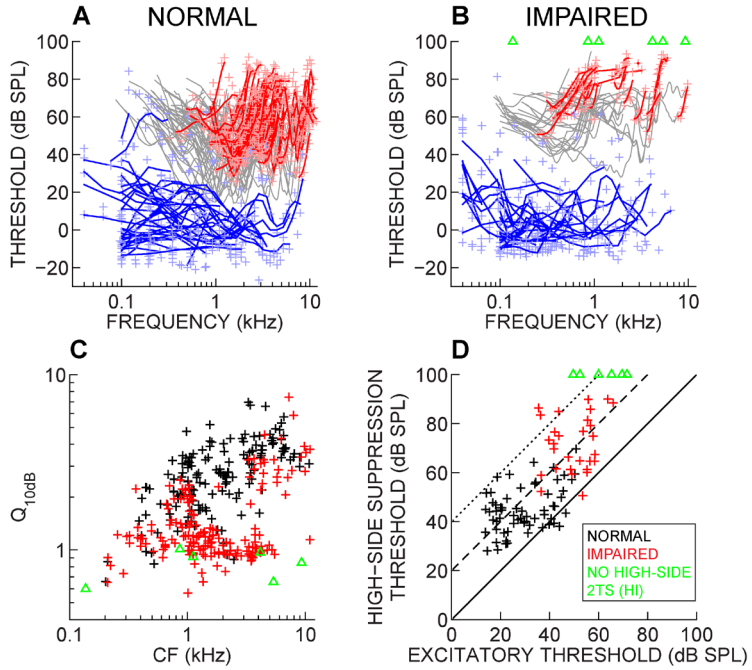


Figure 3. Two-tone-suppression threshold. (A), NH: gray, excitatory-threshold tuning curves; red lines (fits) and crosses (data), high-side suppression-threshold tuning curves; blue lines and crosses, low-side suppression-threshold. (B), HI data. (C), Q_{10dB}. (D), Solid line, equality; dashed and dotted lines, 20- and 40-dB shifts in suppressive threshold, respectively. (B-D), Green triangles at 100-dB SPL, CFs of HI fibers with no high-sided suppression.

3 Results

3.1 Hearing-Loss Characterization

Noise exposure elevated ABR threshold, and reduced DPOAE magnitude, across the audiogram; indicating a mixed inner- and outer-hair-cell pathology (Fig. 2).

3.2 Suppression Threshold

Two-tone-suppression tuning curves were obtained from 62 NH fibers, and 35 HI fibers (Fig. 3A and B). HI excitatory tuning curves show threshold elevation and broadened tuning (Fig. 3C). For NH fibers, high-side 2TS was always observed (Fig. 3A). However, in 6 of 35 HI fibers, we could not detect significant high-side 2TS (Fig. 3B). These fibers have very broadened excitatory tuning (Fig. 3C). In HI fibers with detectable high-side 2TS, the dB difference between excitatory threshold and suppressive threshold was not greater than observed in NH fibers (Fig. 3D). For many fibers high-side-2TS threshold was within 20 dB of on-CF excitatory threshold. Low-side 2TS-threshold estimates are surprisingly low (Fig. 3A and B). All fibers had low-side suppressive regions, regardless of hearing status, typically in the region of 0- to 20-dB SPL.

3.3 Wiener-Kernel Estimates of Suppression

Figure 4 shows second-order Wiener-kernel and STRF analyses from the responses of a single medium-spontaneous-rate (<18 spikes s^{-1}) ANF, with CF 4.2 kHz, recorded in a NH chinchilla. The h_2 kernel is characterized by a series of parallel diagonal lines representing the non-linear interactions driving the ANF's response to noise (Fig. 4A; Recio-Spinoso et al., 2005). The STRF derived from h_2 shows suppressive regions flanking the main excitatory region along the frequency axis, with a high-to-low frequency glide (early-to-late) consistent with travelling-wave delay (Fig. 4B). Decomposing h_2 into its excitatory and suppressive sub-kernels, we observe a broad area of suppressive tuning which overlaps with the excitatory tuning in time and frequency (Fig. 4D and F). There is also an "on-CF" component to the suppression, occurring before the excitation.

There is substantial variability in the relative contribution of suppression to ANF responses (Fig. 5A), similar to that observed with tone stimuli in the "fractional-response" metric (e.g., Miller et al., 1997). There is a group of fibers (both NH and HI) across the CF axis, which show no significant suppression ($R_{\varepsilon,\sigma}$=1; Fig. 5A). These are mainly, but not exclusively, of the high-spontaneous-rate class (≥18 spikes s^{-1}). HI fibers in the region of greatest damage (~2-5 kHz) tend to have reduced suppression. Plotting $R_{\varepsilon,\sigma}$ *vs.* CF-normalized 10-dB bandwidth, we find a significant correlation between broadened tuning and loss of suppression (Fig. 5B). Considering only CFs >2 kHz in this linear regression increased the variance explained from 9.1% to 18.4%.

4 Discussion

Using tones and broadband noise, we found significant changes in the pattern of suppression in the responses of ANFs following noise-induced hearing loss, likely reflecting outer-hair-cell disruption. Previous studies have examined the relationship between ANF 2TS and chronic low-level noise exposure coupled with ageing in the Mongolian gerbil (Schmiedt et al., 1990; Schmiedt & Schultz, 1992), and following an acute intense noise exposure in the cat (Miller et al., 1997). Ageing in a noisy environment was associated with a loss of 2TS. Schmiedt and colleagues often found complete absence of high-side 2TS, with sparing of low-side 2TS, even in cochlear regions with up to 60% outer-hair-cell loss; suggesting potentially different mechanisms for low- and high-side suppression. Our 2TS data are in broad agreement with these earlier findings: HI chinchillas had elevated high-side 2TS thresholds, but retained sensitivity to low-side suppressors. Miller et al. (1997) related a reduction in 2TS in ANFs to changes in the representation of voiced vowel sounds. Weakened compressive nonlinearities contributed to a reduction in "synchrony capture" by stimulus harmonics near vowel-formant peaks. These effects likely contribute to deficits in across-CF spatio-temporal coding of temporal-fine-structure information in speech for HI listeners (e.g., Heinz et al., 2010).

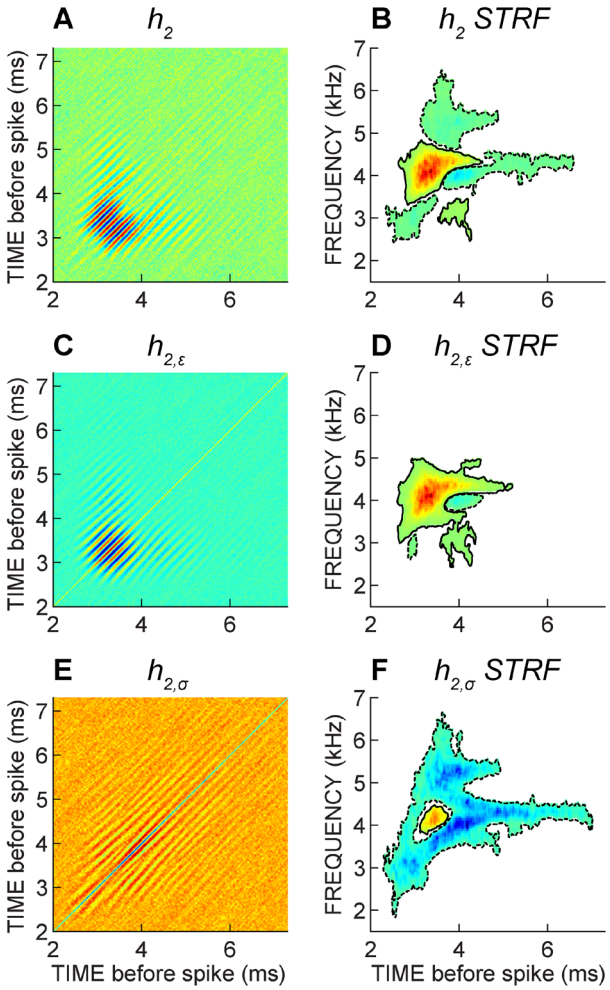


Figure 4. Second-order Wiener-kernel, excitatory and suppressive sub-kernels, and their spectro-temporal receptive-field equivalents calculated from the responses of a single medium-spontaneous-rate ANF in a NH chinchilla: CF = 4.2 kHz, θ = 20-dB SPL, Q_{10} = 3.6, SR = 6.0 spikes s^{-1}. 21,161 driven spike times are included in the analysis. *Right-hand column*, warm colors and solid lines indicate areas of excitation, cool colors and dashed lines areas of suppression, with z-score >3 w.r.t. the bootstrap distribution.

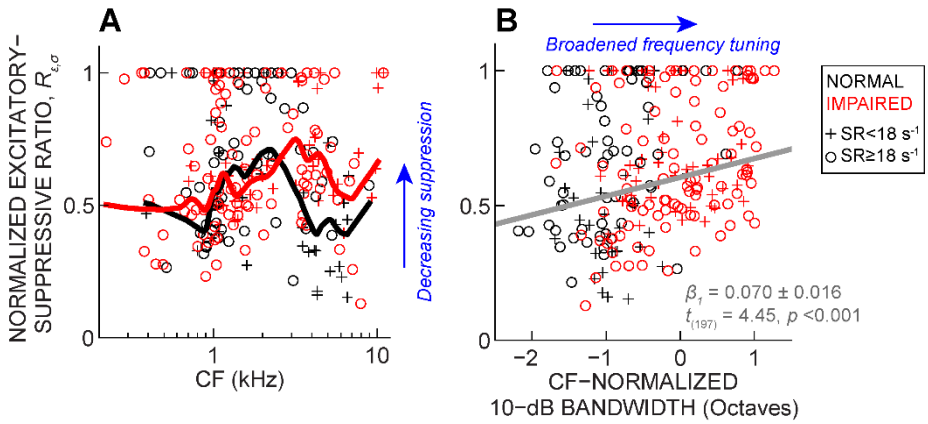


Figure 5. Normalized excitatory-suppressive ratio *vs.* CF (A), and *vs.* CF-normalized 10-dB bandwidth (B). (A), Solid lines, *lowess* fits to values of $R_{\varepsilon,\sigma}$ <1. (B), Horizontal axis expresses 10-dB bandwidth in octaves relative to the 95th percentile of NH-chinchilla ANF data (Kale & Heinz, 2010). Gray line and text, least-squares linear fit to values of $R_{\varepsilon,\sigma}$ <1.

Although the 2TS approach provides important insights on cochlear nonlinearities, it is important to consider the effects of hearing impairment on cochlear nonlinearity in relation to broadband sounds. Our Wiener-kernel approach demonstrated reduced suppression in ANF responses from HI animals, which was correlated with a loss of frequency selectivity. These HI animals also exhibited reduced magnitude DPOAEs: additional evidence of reduced compressive nonlinearity. However, the analyses presented here (Fig. 5) only quantify the overall relative suppressive influence on ANF responses. Using the STRFs derived from h_2s, we aim to characterize the timing and frequency tuning of suppression in the HI auditory periphery in response to broadband sounds. Moreover, these same techniques can be exploited to address potential changes in the balance between excitation and inhibition ("central gain change") in brainstem nuclei following damage to the periphery. Such approaches have previously proved informative across sensory modalities (e.g., Rust et al., 2005), and will likely yield results with tangible translational value.

Acknowledgements

Funded by NIH R01-DC009838 (M.G.H.) and Action on Hearing Loss (UK-US Fulbright Commission scholarship; M.S.). We thank Sushrut Kale, Kenneth S. Henry, and Jon Boley for re-use of ANF noise-response data, and Bertrand Fontaine and Kenneth S. Henry for helpful discussions on spike-triggered neural characterization.

References

Cooper, N. P. (1996). Two-tone suppression in cochlear mechanics. *Journal of the Acoustical Society of America, 99*(5), 3087-3098.

de Boer, E., & Nuttall, A. L. (2002). The mechanical waveform of the basilar membrane. IV. Tone and noise stimuli. *Journal of the Acoustical Society of America, 111*(2), 979-989.

Delgutte, B. (1990). Two-tone rate suppression in auditory-nerve fibers: dependence on suppressor frequency and level. *Hearing Research, 49*(1-3), 225-246.

Geisler, C. D., Yates, G. K., Patuzzi, R. B., & Johnstone, B. M. (1990). Saturation of outer hair cell receptor currents causes two-tone suppression. *Hearing Research, 44*(2-3), 241-256.

Heinz, M. G., Swaminathan, J., Boley, J., & Kale, S. (2010). Across-fiber coding of temporal fine-structure: effects of noise-induced hearing loss on auditory nerve responses. In E. A. Lopez-Poveda, A. R. Palmer & R. Meddis (Eds.), *The Neurophysiological Bases of Auditory Perception* (pp. 621-630). New York: Springer.

Henry, K. S., Kale, S., Scheidt, R. E., & Heinz, M. G. (2011). Auditory brainstem responses predict auditory nerve fiber thresholds and frequency selectivity in hearing impaired chinchillas. *Hearing Research, 280*(1-2), 236-244.

Hicks, M. L., & Bacon, S. P. (1999). Effects of aspirin on psychophysical measures of frequency selectivity, two-tone suppression, and growth of masking. *Journal of the Acoustical Society of America, 106*(3), 1436-1451.

Kale, S., & Heinz, M. G. (2010). Envelope coding in auditory nerve fibers following noise-induced hearing loss. *Journal of the Association for Research in Otolaryngology, 11*(4), 657-673. doi: 10.1007/s10162-010-0223-6

Lewis, E. R., & van Dijk, P. (2004). New variation on the derivation of spectro-temporal receptive fields for primary auditory afferent axons. *Hearing Research, 189*(1-2), 120-136.

Lewis, E. R., Henry, K. R., & Yamada, W. M. (2002a). Tuning and timing in the gerbil ear: Wiener-kernel analysis. *Hearing Research, 174*(1-2), 206-221.

Lewis, E. R., Henry, K. R., & Yamada, W. M. (2002b). Tuning and timing of excitation and inhibition in primary auditory nerve fibers. *Hearing Research, 171*(1-2), 13-31.

Liberman, M. C. (1984). Single-neuron labeling and chronic cochlear pathology. I. Threshold shift and characteristic-frequency shift. *Hearing Research, 16*(1), 33-41.

Miller, R. L., Schilling, J. R., Franck, K. R., & Young, E. D. (1997). Effects of acoustic trauma on the representation of the vowel "eh" in cat auditory nerve fibers. *Journal of the Acoustical Society of America, 101*(6), 3602-3616.

Recio-Spinoso, A., Temchin, A. N., van Dijk, P., Fan, Y. H., & Ruggero, M. A. (2005). Wiener-kernel analysis of responses to noise of chinchilla auditory-nerve fibers. *Journal of Neurophysiology, 93*(6), 3615-3634.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron, 46*(6), 945-956.

Sachs, M. B., & Kiang, N. Y. (1968). Two-tone inhibition in auditory-nerve fibers. *Journal of the Acoustical Society of America, 43*(5), 1120-1128.

Sachs, M. B., & Young, E. D. (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *Journal of the Acoustical Society of America, 68*(3), 858-875.

Schmiedt, R. A., & Schultz, B. A. (1992). Physiologic and histopathologic changes in quiet- and noise-aged gerbil cochleas. In A. L. Dancer, D. Henderson, R. J. Salvi & R. P. Hamernik (Eds.), *Noise-induced hearing loss* (pp. 246-256). St. Louis: Mosby.

Schmiedt, R. A., Mills, J. H., & Adams, J. C. (1990). Tuning and suppression in auditory nerve fibers of aged gerbils raised in quiet or noise. *Hearing Research, 45*(3), 221-236.

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision, 6*(4), 484-507.

Sneary, M. G., & Lewis, E. R. (2007). Tuning properties of turtle auditory nerve fibers: evidence for suppression and adaptation. *Hearing Research, 228*(1-2), 22-30.

Versteegh, C. P., & van der Heijden, M. (2012). Basilar membrane responses to tones and tone complexes: nonlinear effects of stimulus intensity. *Journal of the Association for Research in Otolaryngology, 13*(6), 785-798. doi: 10.1007/s10162-012-0345-0

Versteegh, C. P., & van der Heijden, M. (2013). The spatial buildup of compression and suppression in the mammalian cochlea. *Journal of the Association for Research in Otolaryngology, 14*(4), 523-545. doi: 10.1007/s10162-013-0393-0

Chapter 31

Does Signal Degradation Affect Top-Down Processing of Speech?

Anita Wagner[1,2], Carina Pals[1], Charlotte de Blecourt [2], Anastasios Sarampalis[2,3], Deniz Başkent[1,2]

[1] *University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology, Groningen, Netherlands*

[2] *University of Groningen, Research School of Behavioral and Cognitive Neurosciences, Groningen, Netherlands*

[3] *University of Groningen, Department of Psychology, Groningen, Netherlands*

Contact author: AW, a.wagner@umcg.nl

Abstract

Speech perception is formed based on both the acoustic signal and listeners' knowledge of the world and semantic context. Access to semantic information can facilitate interpretation of degraded speech, such as speech in background noise or the speech signal transmitted via cochlear implants (CIs). This paper focuses on the latter, and investigates the time course of understanding words, and how sentential context reduces listeners' dependency on the acoustic signal for natural and degraded speech via an acoustic CI simulation.

In an eye-tracking experiment we combined recordings of listeners' gaze fixations with pupillometry, to capture effects of semantic information on both the time course and effort of speech processing. Normal-hearing listeners were presented with sentences with or without a semantically constraining verb (e.g., crawl) preceding the target (baby), and their ocular responses were recorded to four pictures, including the target, a phonological (bay) competitor and a semantic (worm) and an unrelated distractor.

The results show that in natural speech, listeners' gazes reflect their uptake of acoustic information, and integration of preceding semantic context. Degradation of the signal leads to a later disambiguation of phonologically similar words, and to a delay in integration of semantic information. Complementary to this, the pupil dilation data show that early semantic integration reduces the effort in disambiguating phonologically similar words. Processing degraded speech comes with increased effort due to the impoverished nature of the signal. Delayed integration of semantic information further constrains listeners' ability to compensate for inaudible signals.

Keywords: Speech perception, degraded speech, cochlear implants

1 Introduction

Processing of speech, especially in one's native language, is supported by world knowledge, the contextual frame of the conversation, and the semantic content. As a consequence, listeners can understand speech even under adverse conditions, where it is partially masked or degraded. Access to these signal-independent sources of information can, however, be compromised if the entire speech signal is degraded, rather than parts of it. This is the case for profoundly hearing impaired listeners who rely on the signal transmitted via a cochlear implant (CI) for verbal

communication. Though CIs allow listeners to perceive speech, this remains an effortful task for them.

In optimal conditions, effortless processing of speech depends on the integration of analyses along a hierarchy of processing stages, as they are described in models of speech perception. These models differ in the way they view the spread of information across various analysis stages (e.g. TRACE: McClelland & Elman, 1986; Shortlist: Norris, 1994; Shorlist B: Norris & McQueen, 2008), but they do agree on the presence of lexical competition. Lexical competition is the process through which listeners consider all the mental representations that overlap with the heard signal as candidates for the word intended by the speaker. Before making a lexical decision listeners thus subconsciously consider multiple words, including homonyms (e.g., 'pair' and 'pear') and lexical embeddings (e.g., *paint* in *paint*ing). In optimal conditions, lexical competition is resolved (i.e. phonologically similar words are disambiguated) very early in the course of speech perception because listeners can rely on a plethora of acoustic cues that mark the difference between phonologically overlapping words (e.g., Salverda, Dahan & McQueen, 2003), and further also benefit from semantic information in sentences (Dahan & Tanenhaus, 2004).

These models are based on data on natural speech perception in optimal conditions, so the question of how analysis of speech is affected by constant degradation of the signal remains unanswered. The present study investigates the time course of lexical competition and semantic integration when processing degraded speech. Furthermore this study will also query whether semantic integration can reduce the mental effort involved in lexical competition in natural and degraded speech. This question has not been studied before since understanding speech in optimal conditions is commonly perceived as effortless. To address these questions we will adapt the approach of Dahan and Tanenhaus (2004), and perform an eye tracking experiment in which listeners are presented with natural and degraded speech. We will further combine the recordings of gaze fixations with pupillometry to obtain a measure of processing effort.

Eye-tracking has been used to study the time course of lexical competition (e.g., Allopena, Magnuson & Tanenhaus, 1998), since listeners' gazes to pictures on the screen reflect their lexical considerations during lexical access as they gradually match the heard signal to an object on the screen. To study the effort involved in processing speech we will record also listeners' change in pupil size. Pupil dilation is a measure that has been used to study effort involved in solving various cognitive tasks (e.g., Hoeks & Levelt, 1993). An increase in pupil dilation has also been shown for listeners presented with degraded speech relative to highly intelligible speech (e.g., Zekveld, et al., 2014). Pupil dilation reflects next to adaptations to changes in luminance or lightness, occurring within the timescale of 200 to 500ms, also a slower evolving response to mental effort, in the timescale of about 900 ms (Hoeks & Levelt, 1993).

2 Methods

2.1 Participants

Twenty-eight native speakers of Dutch, aged between 20 and 30 years (mean = 26), participated in this experiment. None of the participants reported any known hearing or learning difficulties. Their hearing thresholds were normal, i.e. below 20 dB HL on audiometric frequencies from 500 to 8000 kHz. All the participants signed a written consent form for this study as approved by the Medical Ethical Committee of the University Medical Center Groningen. The volunteers received either course credit or a small honorarium for their participation.

2.2 Stimuli

The set of stimuli consisted of the materials used by Dahan and Tananhaus (2004), and an additional set constructed analogously, resulting in a total of 44 critical items. The critical items

were quadruplets of nouns, which were presented together as pictures on the screen. To study the time course of lexical competition we created pairs of critical Dutch words with phonological overlap at the onset, e.g., the target 'pijp' [pipe] was combined with the phonological competitor 'pijl' [arrow]. To study whether disambiguating semantic context reduces lexical competition between acoustically similar words, the two phonologically similar items were presented within sentences, in which a verb that was coherent with only one of these two nouns (e.g. 'rookte' [smoked]) either preceded or followed the noun. The critical pair was presented as pictures together with two Dutch nouns, of which one was semantically viable to follow the verb (e.g., 'kachel' [heater]), the semantic distractor, and the other a phonologically and semantically unrelated distractor ('mossel' [mussel]).

Next to the critical items we constructed 60 sets of filler items. The verbs used in all of these filler sentences were coherent with two nouns, the target and the semantic distractor. The filler items were also presented in quadruplets, and the two remaining distractor nouns were not semantically coherent subjects for the verb. To create a balance between the critical and the filler items, in 20 of the filler items the distractor nouns were phonologically overlapping at the onset. The remaining 40 sets of distractors were phonologically unrelated.

All sentences began with a prepositional phrase, such as "Never before.." or "This morning..". The sentences were recorded from a male native speaker of Dutch. Black and white drawings were created as display pictures, specifically for the purpose of this study.

Two listening conditions were used in the experiment; natural speech (NS) and degraded speech (DS). The degraded stimuli were created using a noise-band-vocoder to simulate CI processing. The stimuli were first bandlimited to 100 to 8000 Hz, and were subsequently bandpass-filtered into 8 channels. Sixth order Butterworth filters were used, with a spacing equal to the distances in the cochlea as determined using the Greenwood function. The slow-varying amplitude envelopes were extracted from each channel via lowpass filtering, and these envelopes were then used to modulate carrier wideband noise, the resulting 8 channels were finally bandpass filtered once more using the same 8 bandpass filters. The processed stimuli were the summed signals from the output of all channels. This manipulation lead to stimuli that preserved the durational cues in an otherwise unnatural and spectrotemporally degraded acoustic form.

2.3 Procedure

Before data collection, participants were familiarized with the pictures and the nouns that refer to the pictures. They were then seated in a comfortable chair facing the monitor, and an Eyelink 500 eye-tracker was mounted and calibrated. This head mounted eye-tracker contains two small cameras, which can be aligned with the participants' pupil to track the pupil's movements and size continuously during the experiment. Pupil size was recorded together with gaze fixations using a sampling rate of 250 Hz.

The stimuli were presented via a speaker in sound attenuated room. The lighting in this room was kept constant throughout the experiment to avoid effects of ambient light intensity on the pupil diameter. The participants' task was to listen to the stimuli and to click on the picture corresponding to the target noun in the sentence. Each participant was presented with stimuli blocked into an NS and DS condition. Before the DS condition, the participants were familiarized with the degradation used in this study by listening to degraded sentences and selecting the correct sentence from 10 sentences presented on the screen.

Each experimental item was presented only once in either the context or neutral sentence, and in either NS or DS. Between the two blocks (NS and DS) there was a break. Four practice trials preceded each block (using filler items), and a block consisted of 48 experimental items; 22 critical items and 26 filler items. The order of the presentation between blocks and items was quasi-random.

2.4 Analysis

Trials in which participants clicked on the wrong picture were excluded from the analysis. Trials with eye blinks longer than 300 ms were also excluded. Shorter blinks were corrected for by means of linear interpolation.

Gaze Fixations

To address the question of how semantic context affects lexical competition between phonologically similar words the statistical analyses focus on listeners' gaze fixations towards the phonological competitor and the semantic distractor. The probabilities of gaze fixations towards this competitor and this distractor were statically analyzed by means of growth curves (Mirman, 2014). R (R Core team, 2013) with lme4 package (Bates et al., 2014) was used to model the time curves of fixations as 4th order polynomials within the time window of 200-2000 ms after word onset. Two logistic-regression multi-level models were used, with fixations to either the phonological competitor or the semantic distractor, coded as a binomial response. The time course curves were described in four terms: intercept, the overall slope of the curve, the width of the rise and fall around the inflection, and the curvature in the tails. The probability of fixations along the time course was modeled as a function of Context (neutral versus context), Presentation (NS versus DS) and the possible three-way interactions between these two factors and all four terms describing the curves. As random effect, we included individual variation among participants and items on all four terms describing the time curve. Model comparison was used to estimate the contribution of individual predictors to the fit of the model. For this, individual fixed effects were sequentially added, and the change in the model fit was evaluated by means of likelihood ratio test.

Pupil dilation

To investigate the effort involved in the process of lexical competition with and without semantic context, the pupil dilation data per participant were baseline-corrected to the 200 ms preceding the presentation of the experimental item. The baseline-corrected data were normalized to correct for individual differences in pupil size, according to the equation:

$$\%\ Event\ Related\ Pupil\ Dilation = \frac{observation - baseline}{baseline} \times 100.$$

For the statistical analysis, the pupil size data, as captured by the event-related pupil dilation (ERPD), were analyzed analogously to the fixation data, as time curves of pupil dilation. The time-course functions were analyzed as 3rd-order polynomials, since, during fitting, the fourth order turned out to be redundant to the description of these curve functions. The terms describing the curves are: intercept, the slope of the curve, and a coefficient for the curvature around the inflection point. These time curves were analyzed by means of multi-level nonlinear regression model. The statistical models contained in addition to the terms describing the curves per participant also random effects on these three terms per participant, and for the phonological competitor model also random effects per item.

3 Results

3.1 Gaze fixations

Figure 1 displays the time curves of fixations to all four pictures displayed within the NS blocks for C (a), and N (b), and for the DS blocks for C (c) and N (d). These figures show proportions of fixations to the four pictures displayed, averaged across participants, and the 95% confidence intervals for the fixations to the target and competitor.

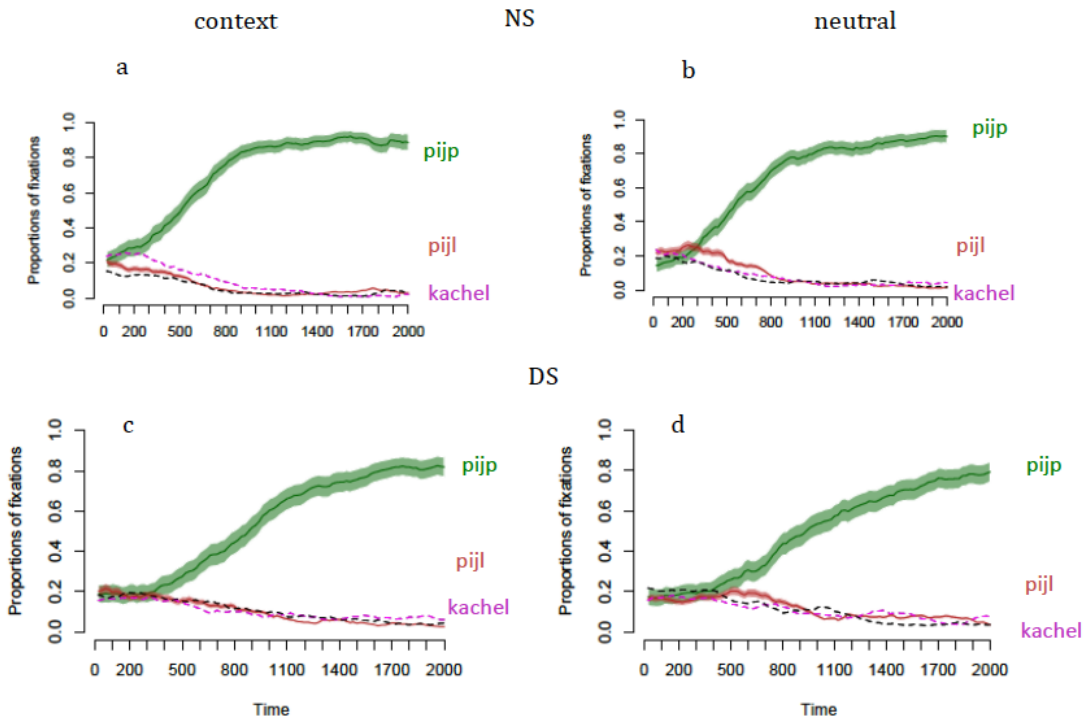


Figure 1. Time curves of fixations to the pictures displayed for NS (a & b) and DS (c & d), and for items presented in context sentences (a & c), and neutral sentences (b & d).

Of particular interest for this study are the three-way interactions between Context (C versus N) and Presentation (NS versus DS) and the terms describing the course of the curves. For the fixations to the phonological competitor, as significant emerged the three way interactions with the intercept of the curve ($\chi^2(18)= 28476$, $p<0.001$, the interaction with the quadratic term (the slope), ($\chi^2(18)= 28184$, $p<0.001$), the interaction between the cubic term (rise and fall around the central inflection), ($\chi^2(18)= 27632$, $p<0.001$), and the quartic term (curvature in the tails), ($\chi^2(18)= 27651$, $p<0.05$). The interaction with the intercept shows that the context sentences reduced the area under the fixation curves to the competitor for NS (red lines in Figure 1 a versus b), and that this reduction was smaller for DS (red lines in Figure 1 c versus d). The interaction with the slope shows that the growth of fixations to the competitor is shallower for DS in the neutral context than it is for NS in neutral context. The interaction with the cubic term reflects that the location of the peak of fixations towards the competitor in DS is delayed for about 300 ms relative to the location of the peak for NS, and that the course of this curve is more symmetric than for NS, and this mainly for the items presented in neutral context. The interaction with the quartic term reflects a slower decline of fixations towards the competitor in DS versus NS, and shallower for items in context than in neutral sentences.

For the fixations to the semantic distractor, as significant emerged the interactions between Context and Presentation and the intercept of the curve ($\chi^2(3)= 2268.6$ $p<0.001$), the interaction with the quadratic term ($\chi^2(3)= 337.25$, $p<0.001$), the interaction between the cubic term, ($\chi^2(3)= 69.41$, $p<0.001$), and the quartic term ($\chi^2(3)= 19.09$, $p<0.05$). These interactions reflect what can also be seen in a comparison of between NS and DS in Figure 1. Namely that in NS, listeners fixate the semantic competitor more often in the context sentences than in neutral context. This effect is absent for DS.

3.2 Pupil dilation

Figure 2 displays the time course of pupil dilation for NS and DS and for the two contexts C and N.

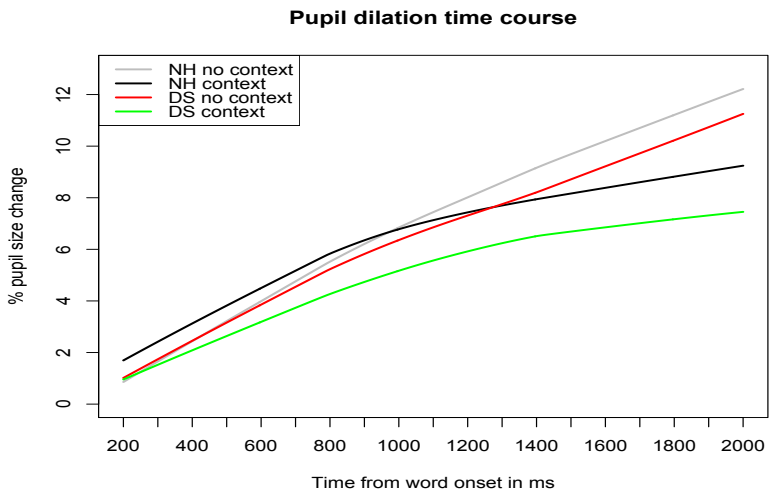


Figure 2. Time curves of pupil dilation, averaged across participants for NS (black and grey) and DS (red and green), and for items presented in context sentences (black and green), and neutral sentences (red and grey).

The curves for NS and DS in the neutral condition show a constant increase in pupil size over time as a function of lexical competition. The curves for the context condition show a decline in pupil size growth starting at around 800 ms after the onset of the target word. The statistical analysis revealed significant three way interactions with Context (N versus C) and Presentation (NS versus DS) on all terms describing the curves: Intercept ($\chi^2(3)= 301.90$, $p < .001$), slope ($\chi^2(3)= 145.3$, $p < .001$), and the cubic term, the curvature around the peak ($\chi^2(3)= 272.52$, $p < .001$). This implies that pupil dilation was sensitive in capturing the reduced effect of lexical competition in the context sentences versus neutral context, but this effect was delayed and smaller in DS than in NS.

A look at this figure suggests that the effort involved in lexical competition for DS was overall smaller for DS than for NS. This overall smaller increase in pupil dilation can be explained by the fact that these curves are normalized to a baseline of 200 ms preceding the presentation of each critical item per trial, participant and condition. Listening to degraded speech is by itself more effortful than listening to natural speech (e.g., Winn, Edwards & Litovsky, 2015), and therefore there is a difference in the baseline between DS and NS. These differences in the baseline can be explained by the difference in processing degraded versus natural speech, and are independent of the effects of semantic integration on lexical competition.

4 Discussion

This present study examined the effect of semantic integration on the time course of lexical competition, and on the effort involved in solving lexical competition in natural and degraded speech. Our results show that processing natural speech comes with a timely integration of semantic information, which in turn reduces lexical competition. Listeners are then able to pre-select a displayed target based on its semantic coherence with the context, and this allows listeners to reduce the effort involved in lexical competition. When processing degraded speech the integration of semantic information is delayed, as is also lexical competition. This implies that semantic integration is not able to reduce lexical competition, which by itself is longer and occurs later. These results were also mirrored by the pupil dilation data, in which a release from lexical competition was visible but delayed. Mapping of degraded speech to mental representations is more effortful due the mismatch between the actual signal and its mental representation, and lexical context is not able to release listeners from this effort on time. In natural situations, in which words are being heard in succession, and the speech signal evolves

quickly over time, such a difference in processing speed of degraded speech will accumulate effort, and draw more strongly on resources in working memory.

Acknowledgments

We would like to thank Dr. Paolo Toffanin for technical support, and Prof. Frans Cornelissen (University Medical Center Groningen) for providing the eye-tracker for this study. This work was supported by a Marie Curie Intra-European Fellowship (FP7-PEOPLE-2012-IEF 332402). Support for the second author came from a VIDI Grant from the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw) Grant No. 016.093.397. The study is part of the research program of our department: Healthy Aging and Communication.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38, 419–439

Bates D., Maechler M., Bolker B. and Walker S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, http://CRAN.R-project.org/package=lme4

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 498-513

Hoeks B., Levelt W. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. Behavior Research Methods, 25(1), 16–26

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, l-86

Mirman, D. (2014). Growth Curve Analysis and Visualization Using R. Chapman and Hall / CRC

Norris, D. (1994) Shortlist: A connectionist model of continuous speech recognition. Cognition, 52, 189-234

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review, 115(2), 357-395

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Salverda, A.P., Dahan, D., & McQueen, J.M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition, 90, 51-89

Winn, M.B., Edwards, J.R.., and Litovsky, R.Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. Ear and Hearing (ahead of print)

Zekveld, A.A., Heslenfeld, D.J., Johnsrude, I.S., Versfeld, N., & Kramer, S.E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. NeuroImage, 101, 76–86

Chapter 32

The Effect of Peripheral Compression on Syllable Perception measured with a Hearing Impairment Simulator

Toshie Matsui, Toshio Irino, Misaki Nagae, Hideki Kawahara
Faculty of Systems Engineering, Wakayama University, Japan

Roy D. Patterson
Department of Physiology, Development and Neuroscience, Centre for the Neural Basis of Hearing, University of Cambridge, UK

Contact author: TM, tmatsui@sys.wakayama-u.ac.jp

Abstract

Hearing impaired (HI) people often have difficulty understanding speech in multi-speaker or noisy environments. With HI listeners, however, it is often difficult to specify which stage, or stages, of auditory processing are responsible for the deficit. There might also be cognitive problems associated with age. In this paper, a HI simulator, based on the dynamic, compressive gammachirp (dcGC) filterbank, was used to measure the effect of a loss of compression on syllable recognition. The HI simulator can counteract the cochlear compression in normal hearing (NH) listeners and, thereby, isolate the deficit associated with a loss of compression in speech perception. Listeners were required to identify the second syllable in a three-syllable "nonsense word", and between trials, the relative level of the second syllable was varied, or the level of the entire sequence was varied. The difference between the Speech Reception Threshold (SRT) in these two conditions reveals the effect of compression on speech perception. The HI simulator adjusted a NH listener's compression to that of the "average 80-year old" with either normal compression or complete loss of compression. A reference condition was included where the HI simulator applied a simple 30-dB reduction in stimulus level. The results show that the loss of compression has its largest effect on recognition when the second syllable is attenuated relative to the first and third syllables. This is probably because the internal level of the second syllable is attenuated proportionately more when there is a loss of compression.

Keywords: hearing impairment simulator, cochlear compression, dynamic compressive gammachirp filterbank, audiogram

1 Introduction

Age related hearing loss (presbycusis) makes it difficult to understand speech in noisy environments and multi-speaker environments (Moore, 1995; Humes, 2010). There are several factors involved in presbycusis: loss of frequency selectivity, recruitment, and loss of temporal fine structure (Moore, 2007). The isolation and measurement of these different aspects of hearing impairment (HI) in elderly listeners is not straightforward, partly because they often have multiple auditory problems, and partly because they may also have more central cognitive

problems including depression. Together the problems make it difficult to obtain sufficient behavioral data to isolate different aspects of HI (Jerger et al., 1989; Lopez et al., 2003). We have developed a computational model of HI and a HI simulator that allow us to perform behavioural experiments on cochlear compression – experiments in which young normal hearing (NH) listeners act as "patients" with isolated, sensory-neural hearing losses of varying degree.

The HI simulator is essentially a high-fidelity vocoder based on the dynamic, compressive gammachirp filter bank (dcGC-FB) (Irino & Patterson, 2006). It analyses naturally recorded speech sounds with a model of the auditory periphery that includes fast-acting compression and then resynthesizes the sounds in a form that counteracts the compression in a NH listener. They hear speech with little distortion or background noise (Irino et al., 2013) but without their normal compression. This makes it possible to investigate recruitment phenomenon that are closely related to loss of cochlear compression (Bacon et al., 2004) using NH listeners who have no other auditory problems. Specifically, we describe an experiment designed to examine how the loss of cochlear compression affects the recognition of relatively soft syllables occurring in the presence of louder, flanking syllables of varying levels. The aim is to reveal the role of cochlear compression in multi-speaker environments or those with disruptive background noise.

2 Method

Listeners were presented three-syllable "nonsense words" and required to identify the second syllable. Between trails, the level of the second syllable within the word was varied ("dip" condition) or the level of the entire word was varied ("constant" condition). The difference between second-syllable recognition in the two conditions reveals the effect of compression. The HI simulator is used to adjust syllable level to simulate the hearing of 80 year old listeners with "normal" presbycusis (80yr 100%) or a complete loss of compression (80yr 0%).

2.1 Participants

Ten listeners participated in the experiment (four males; average age: 23.6 years; SD: 4.77 years). None of the listeners reported any history of hearing impairment. The experiment was approved by the ethics committee of Wakayama University and all listeners provided informed consent. Participants were paid for their participation, except for the first and third authors who also participated in the experiment.

2.2 Stimuli and Procedure

The stimuli were generated in two stages: First, three-syllable nonsense words were composed from the recordings of a male speaker, identified as "MIS" (FW03) in a speech sound database (Amano et al., 2006), and adjusted in level to produce the dip and constant conditions of the experiment. Second, the HI simulator was used to modify the stimuli to simulate the hearing of two 80yr old listeners, one with average hearing for an 80yr old and one with a complete loss of compression. There was also a normal-hearing control condition in which the level was reduced a fixed 30 dB.

In the first stage, nonsense words were composed by choosing the *second* syllable at random from the 50 Japanese syllables presented without parentheses in Figure 1. This is the list of syllables (57-S) recommended by the Japanese Audiological Society (2003) for use in clinical studies to facilitate comparison of behavioural data across studies. The first and third syllables in the words were then selected at random from the full set of 62 syllables in Figure 1. The set was enlarged to reduce the chance of listeners recognizing the restriction on the set used for the target syllable. Syllable duration ranged from 211 ms to 382 ms. There was a 50-ms silence between adjacent syllables within a word.

a	ka	sa	ta	na	ha	ma	ja	ɾa	wa	ga	(dza)	da	ba
i	kʲi	ʃi	tʃi	ɲi	çi	mi		ɾʲi		(gʲi)	dʒi		(bʲi)
ɯ	kɯ	sɯ	tsɯ	(nɯ)	ɸu	mɯ	jɯ	ɾɯ		(gɯ)	dzɯ		(bɯ)
e	ke	se	te	ne	he	me		(ɾe)		(ge)	(dze)	de	(be)
o	ko	so	to	no	ho	mo	jo	ɾo		go	(dzo)	do	(bo)

Figure 1. The 62 Japanese syllables used in the experiment using the international phonetic alphabet. The syllables in parentheses are also Japanese syllables but not included in list 57-S. All 62 were used in the draw for the first and third syllables. The draw for the second syllable was restricted to the 50 syllables not enclosed in parentheses.

The nonsense words were then adjusted to produce the two level conditions illustrated in Figure 2: In the "dip" condition (left column), the level of the second syllable was 70, 60, 50 or 40 dB, while the first and third syllables were fixed at 70 dB. In the "constant" condition (right column), all three syllables were set to 70, 60, 50 or 40 dB.

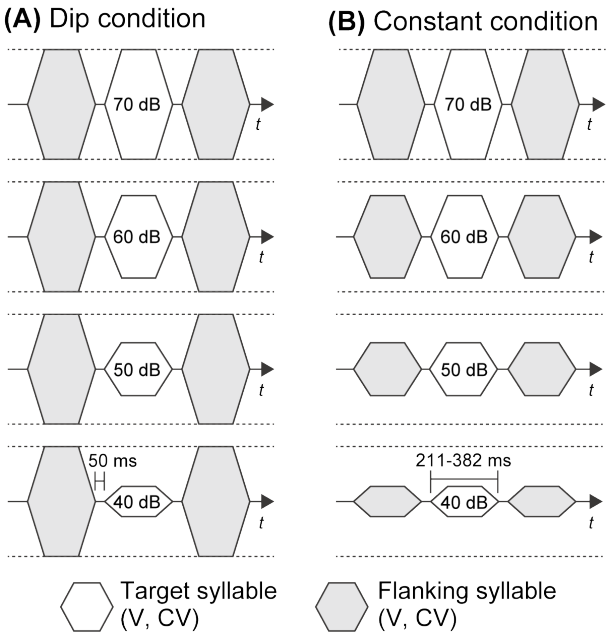


Figure 2. Schematic of the experimental paradigm. The levels of the nonsense words were adjusted to produce stimuli for (A) the Dip condition and (B) the Constant condition. The 70-dB stimuli were the same in the two conditions.

In the second stage, the stimuli were processed to simulate:

1) the average hearing level of an 80 year old person (audiogram ISO7029, ISO/TC43, 2000) simulated with 100% compression (80yr 100%),

2) the average hearing level of an 80-year-old simulated with 0% compression (80yr 0%),

3) normal adult hearing with a 30-dB reduction across the entire frequency range (NH - 30dB).

To realize the 80yr 100% condition, the output level of each channel was decreased until the audiogram of the dcGC-FB (normal hearing) matched the average audiogram of an 80yr old. To realize the 80yr 0% condition, an audiogram was derived with dcGC-FB having no compression whatsoever, and then the output level of each channel was decreased until the audiogram matched the audiogram of an average 80yr old. In the compression-cancellation process, loss of hearing level was limited so that the audiogram did not fall below the hearing level of an

average 80yr old. The compression was applied using the signal processing system described in Nagae et al. (2014).

Each participant performed all conditions – a total of 1050 trials or nonsense words: 50 words x 7 dip/constant conditions x 3 hearing impairment simulations. The participants were required to identify the second syllable using a GUI on a computer screen. The stimuli were manipulated, and the results collected using MATLAB. The sounds were presented dichotically via a DA converter (Fostex, HP-A8) over headphones (Sennheiser, HD-580). The experiment was carried out in a sound-attenuated room with a background level of about 26 dB L_{Aeq}.

3 Results

For each participant, the four percent-correct values for each level condition (dip or constant) were fitted with a cumulative Gaussian psychometric function to provide a Speech Reception Threshold (SRT) value for that condition, separately in each of the three simulated hearing-impairment conditions. The cumulative Gaussian was fitted by the bootstrap method (Wichmann & Hill, 2011a; Wichmann & Hill, 2011b) and SRT was taken to be the sound pressure level associated with 50 percent correct recognition of the second syllable. The SRT in the constant condition with the NH -30 dB hearing impairment was taken as standard performance on the task. The difference between this standard SRT and the SRTs of the other conditions are the experimental results (ΔSRTs) for each participant. Figure 3(A) shows the mean and standard deviation over participants of ΔSRT for each condition.

A two-way ANOVA was performed for the dip/constant conditions and the hearing impairment conditions combined, and it showed that the manipulations had a significant effect on recognition performance (ΔSRT) ($F(5, 54) = 23.82$, $p < 0.0001$). Significant, separate main effects were also observed for the dip/constant condition and the hearing impairment ($F(1, 54) = 41.99$, $p < 0.0001$; $F(2, 54) = 36.11$, $p < 0.0001$, respectively). The interaction between the two factors was not significant ($\alpha = 0.05$). Multiple comparisons with the Tukey-Kramer HSD test were performed. There were significant differences between all of the hearing impairment conditions ($\alpha = 0.05$).

ΔSRT provides a measure of the difficulty of the task, and the statistical analysis shows that the different hearing impairment simulations vary in their difficulty. In addition, ΔSRT for the dip conditions is larger than ΔSRT for constant conditions for all of the hearing impairment simulations. This difference in difficulty could be the result of forward masking since the gap between offset of the first syllable and the onset of the target syllable is less than 100 ms, and the third syllable limits any post processing of the second syllable. The dip condition would be expected to produce more masking than the constant condition, especially when the target syllable is sandwiched between syllables whose levels are 20 dB or 30 dB greater. The constant condition only reveals differences between the different hearing impairment simulations. This does, however, include differences in audibility.

Figure 3(B) shows the difference between SRT in the dip and constant conditions; that is, ΔSRT relative to the constant condition. It provides a measure of the amount of any forward masking. A one-way ANOVA was performed to confirm the effect of hearing impairment simulation ($F(2, 29) = 6.14$, $p = 0.0063$). A multiple comparisons test (Tukey-Kramer HSD) showed that ΔSRT in the 80yr 0% condition was significantly greater than in the 80yr 0% condition, or the NH -30 dB condition ($\alpha = 0.05$).

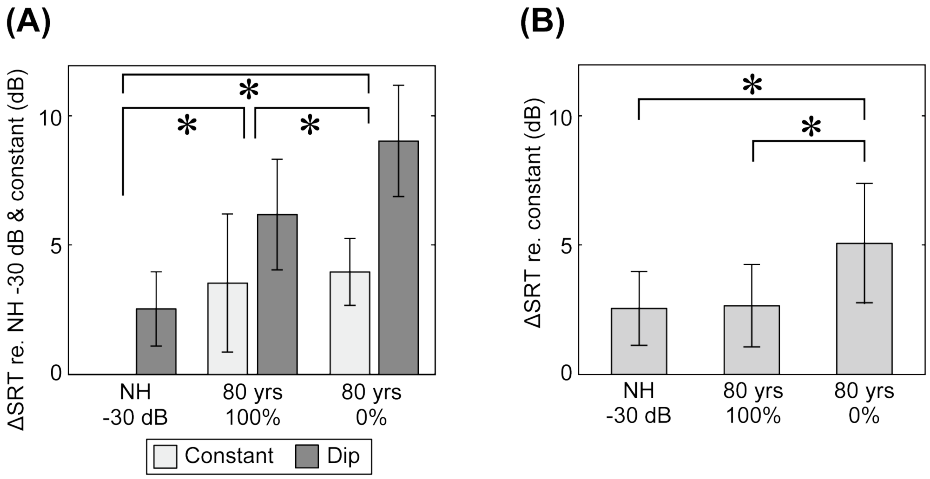


Figure 3. (A) The difference in SRT (ΔSRT) between the constant condition of the NH -30 dB simulation and the SRTs of the other conditions. (B) The SRT difference between the dip and constant conditions for each hearing impairment condition. Asterisks (*) show significant differences ($\alpha = 0.05$) in a multiple comparison, Tukey-Kramer HSD test. Error bars show the standard deviation.

4 Discussion

It is assumed that forward masking occurs during auditory neural processing, and that it operates on peripheral auditory output (i.e. neural firing in the auditory nerve). Since the compression component of the peripheral processing is normal in both the 80yr 100% condition and the NH -30dB condition, the level difference between syllable 2 and syllables 1 and 3 should be comparable in the peripheral output of any specific condition. In this case, ΔSRT provides a measure of the more central masking that occurs during neural processing. When the compression is cancelled, the level difference in the peripheral output for syllable 2 is greater in the dip conditions. If the central masking is constant, independent of the auditory representation, the reduction of the output level for syllable 2 should be reflected in the amount of masking, i.e. ΔSRT. This is like the description of the growth of masking by Moore (2012), where the growth of masking in a compressive system is observed to be slower than in a non-compressive system. Moore was describing experiments where both the signal and the masker were isolated sinusoids. The results of the current study show that the effect generalizes to speech perception where both the target and the interference are broadband, time-vary sounds.

There are, of course, other factors that might have affected recognition performance: The fact that the same speaker provided all of the experimental stimuli might have reduced the distinctiveness of the target syllable somewhat, especially since syllable 2 was presented at a lower stimulus level in the dip conditions. The fact that listeners had to type in their responses might also have decreased recognition performance a little compared to everyday listening. It is also the case that our randomly generated nonsense words might accidentally produce a syllable sequence that is similar to a real word with the result that the participant answers with reference to their mental dictionary and makes an error. But, all of these potential difficulties apply equally to all of the HI conditions, so it is unlikely that they would affect the results.

This suggests that the results truly reflect the role of the loss of compression when listening to speech in everyday environments for people with presbycusis or other forms of hearing impairment. It shows that this new, model-based approach to the simulation of hearing impairment can be used to isolate the compression component of auditory signal processing. It may also assist with isolation and analysis of other aspects of hearing impairment.

5 Summary

This paper reports an experiment performed to reveal the effects of a loss of compression on the recognition of syllables presented with fluctuating level. The procedure involves using a filterbank with dynamic compression to simulate hearing impairment in normal hearing listeners. The results indicate that, when the compression of normal hearing has been cancelled, it is difficult to hear syllables at low sound when they are sandwiched between syllables with higher levels. The hearing losses observed in patients will commonly involve multiple auditory problems and they will interact in complicated ways that make it difficult to distinguished the effects of any one component of auditory processing. The HI simulator shows how a single factor can be isolated using an auditory model, provided the model is sufficiently detailed.

Acknowledgments

This work was supported in part by JSPS Grants-in-Aid for Scientific Research (Nos 24300073, 24343070, 25280063 and 25462652).

References

Amano, S., Sakamoto, S., Kondo T., & Suzuki, Y. (2006). NTT-Tohoku University familiarity controlled word lists (FW03). Speech resources consortium, National Institute of Informatics., Japan.

Bacon, S. P., Fay, R., & Popper, A. N. (Eds.) (2004). *Compression: from cochlea to cochlear implants*. New York: Springer.

Humes, L. E., & Dubno, J. R. (2010). Factors affecting speech understanding in older adults. In S. Gordon-Salant, R. D. Frisina, A. N. Popper, & R. R. Fay (Ed.), *The aging auditory system* (pp. 211-257). New York: Springer.

Irino, T., & Patterson, R. D. (2006). A dynamic compressive gammachirp auditory filterbank. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6), 2222—2232.

Irino, T., Fukawatase, T., Sakaguchi, M., Nisimura, R., Kawahara, H., & Patterson, R. D. (2013). Accurate estimation of compression in simultaneous masking enables the simulation of hearing impairment for normal hearing listeners. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, H. E. Gockel (Eds.), *Basic Aspects of Hearing, Physiology and Perception* (pp.73—80). New York: Springer.

ISO/TC43 (2000). Acoustics - Statistical distribution of hearing thresholds as a function of age. Geneva: International Organization for Standardization.

Japan Audiological Society (2003). The 57-S syllable list. *Audiology Japan*, 46, 622—637.

Jerger, J., Jerger, S., Oliver, T., & Pirozzolo, F. (1989). Speech understanding in the elderly. *Ear and Hearing*, 10(2), 79—89.

Lopez, O. L., Jagust, W. J., DeKosky, S. T., Becker, J. T., Fitzpatrick, A., Dulberg, C., Breitner, J., Lyketsos, C., Jones, B., Kawas, C., Carlson, M., & Kuller, L. H. (2003). Prevalence and classification of mild cognitive impairment in the cardiovascular health study cognition study: part 1. *Archives of neurology*, 60(10), 1385—1389.

Moore, B. C. J. (1995). *Perceptual Consequences of Cochlear Damage*. New York: Oxford University Press.

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. Chichester: John Wiley & Sons.

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing* (6th ed.). Bingley: Emerald Group Publishing Limited.

Nagae, M., Irino, T., Nishimura R., Kawahara, H., & Patterson, R. D. (2014). Hearing impairment simulator based on compressive gammachirp filter. *Proceedings of APSIPA ASC*.

Wichmann, F. A., & Hill, N. J. (2011). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8), 1314-1329.

Wichmann, F. A., & Hill, N. J. (2011). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & psychophysics*, 63(8), 1293-1313.

Chapter 33

Towards Objective Measures of Functional Hearing Abilities

Hamish Innes-Brown[1,2], Renee Tsongas[1], Jeremy Marozeau[1,3], Colette McKay[1]

[1] *Bionics Institute, Melbourne, Australia*
[2] *Research Group Experimental Oto-laryngology, KU Leuven, Belgium*
[3] *Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark*

Contact author: HIB, hinnes-brown@bionicsinstitute.org

Abstract

Aims: People with impaired hearing often have difficulties in hearing sounds in a noisy background. This ability relies partly on the capacity of the auditory system to process temporal information in the sound signal. In this study we examined the relationships between perceptual sensitivity to temporal fine-structure cues, brainstem encoding of complex harmonic and amplitude-modulated sounds, and the ability to understand speech in noise. Understanding these links will allow the development of an objective measure that could be used to detect changes in functional hearing before the onset of permanent threshold shifts.

Methods: We measured TFS sensitivity and speech in noise performance (QuickSIN) behaviourally in 34 normally-hearing adults with ages ranging from 18 to 34 years. We recorded brainstem responses to complex harmonic sounds and a 4000 Hz carrier signal modulated at 110 Hz. We performed cross-correlations between the stimulus waveforms and scalp-recorded brainstem responses to generate a simple measure of stimulus encoding accuracy, and correlated these measures with age, TFS sensitivity and speech-in-noise performance.

Results: Speech-in-noise performance was negatively correlated with TFS sensitivity and age. TFS sensitivity was also positively correlated with stimulus encoding accuracy for the complex harmonic stimulus, while increasing age was associated with lower stimulus encoding accuracy for the modulated tone stimulus.

Conclusions: The results show that even in a group of people with normal hearing, increasing age was associated with reduced speech understanding, reduced TFS sensitivity, and reduced stimulus encoding accuracy (for the modulated tone stimulus). People with good TFS sensitivity also generally had more faithful brainstem encoding of a complex harmonic tone.

Keywords: Speech in noise, objective measures, temporal fine-structure

1 Introduction

The benefits of amplification are greatest when hearing interventions are made as early as possible. There is therefore great interest in the clinical audiology community in the development of objective techniques to measure various hearing abilities that do not require behavioural responses from the patient and are able to determine fitting parameters for HAs (and cochlear implants). While the use of cortical responses is in development (Billings et al., 2007; Carter et al., 2010; Billings et al., 2011; Billings, 2012; Chang et al., 2012), transient

brainstem responses (Audiology, 2003) and steady state responses (Luts et al., 2004; Luts et al., 2006; Alaerts et al., 2010) are already used in clinical practice to objectively assess hearing thresholds in young children. However, while these measures provide an estimate of *audibility* with the prescribed HA gain, they do not provide any indication of the expected *hearing ability* of the patient.

In order to develop an objective measure of hearing ability, it is necessary to establish the links between speech perception, psychophysical measures of perceptual sensitivity to the acoustic cues that underlie effective speech perception, and the proposed objective measures. This paper describes our initial investigations in this direction.

The overall aim of this study was to examine the relationships between perceptual sensitivity to temporal fine-structure cues, brainstem encoding of complex harmonic and amplitude-modulated sounds, and the ability to understand speech in noise. Understanding these links will allow the development of an objective measure that could be used to detect changes in functional hearing before the onset of permanent threshold shifts.

2 Methods

2.1 Participants

34 participants (14 men and 20 women) aged between 18-63 years took part in the experiment. All participants had normal hearing bilaterally, defined by four frequency average hearing loss thresholds (500Hz, 1kHz, 2kHz and 4kHz) of less than 25 dB HL. Pure tone hearing thresholds for the 34 participants are shown in Figure 1. Thresholds were measured using an Otometrics MADSEN Itera II audiometer with TDH-39 Stereo Headphones. The study was approved by RVEEH Human Research Ethics Committee. Consent was written and informed.

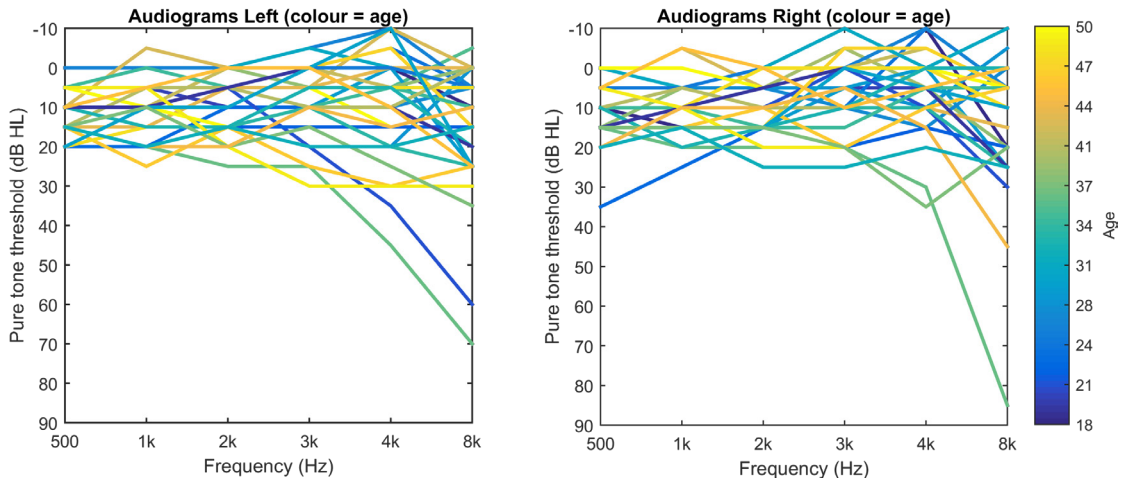


Figure 1. Pure tone hearing thresholds for each ear.

2.2 Temporal fine structure sensitivity

The TFS1 test (Moore & Sek, 2009, 2012) was used to measure participants' temporal fine structure (TFS) sensitivity. The test was carried out using the TFS1 test software (http://hearing.psychol.cam.ac.uk/) and was based on the standard TFS1 test protocol. The task was performed on the participant's better ear, as determined by the audiogram (4FAHL).

One practice run was given prior to the test. If participants could perform the task and attain a threshold they were given three real runs. For some participants the staircase procedure saturated (reached the 'easy' limit). Instead of receiving a threshold, these participants received

a percent-correct score obtained from 40 trials of the 2AFC task at the easiest level of the staircase. Both the threshold and the percent-correct scores were converted to a d' sensitivity measure using the method outlined by Hopkins & Moore (2007).

2.3 Speech in noise tests (QuickSIN)

Speech-in-noise performance was assessed behaviourally using the QuickSIN test (Killion et al., 2004). Six sets of six sentences with five key words per sentence were presented to both ears in four-talker babble noise. The sentences were presented with decreasing signal-to-noise ratios (SNRs) from +25 dB SNR to 0 dB SNR in 5 dB steps (+25, +20, +15, +10, +5 and 0 dB). The average SNR loss was calculated across the six lists. This score indicates the increase in SNR required for the participant to understand the sentence compared with a normal hearing person. Any score less than 4 dB is considered normal, and a lower SNR loss score reflects better speech-in-noise performance.

2.4 Electrophysiology

Stimuli

Envelope-following responses (EFRs) from the brainstem were elicited in response to two stimulus types – a complex harmonic tone, and a 4 KHz carrier tone modulated at 110 Hz. Both sounds were 100 ms in duration with 5 ms onset and offset ramps. The complex harmonic tone had an F0 of 180 Hz and 20 equally-spaced harmonics of equal amplitude and random phase. The modulated tone had a modulation depth of 50 % (-6 dB). Each stimulus was presented with alternating polarities. The first 30 ms of each stimulus are shown in Figure 2.

Stimuli were controlled via custom software in MAX/MSP (Cycling '74), played through an RME Fireface 400 audio interface and Etymotic ER3-A insert-phones.

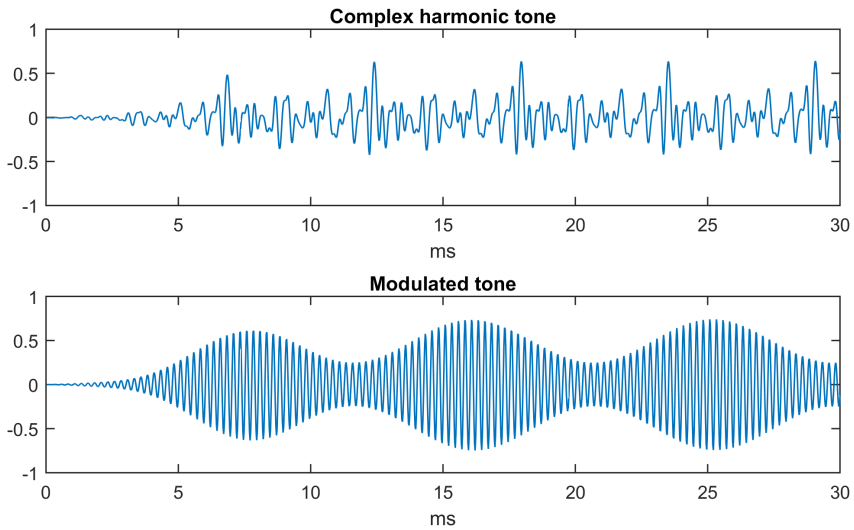


Figure 2. The first 30 ms of the complex harmonic tone (top) and modulated tone (bottom).

EEG Recordings and Pre-processing

EEG data was recorded from the scalp via Ag-AgCl electrodes, using a BioSemi ActiveTwo EEG System and ActiView software. Electrode offsets were ± 40mV. The EEG data was collected in continuous mode at a sampling rate of 16.384 kHz. The -3 dB point of the antialiasing lowpass filter was 3334 Hz.

The EEG recordings were segmented into epochs of -50 to 150 ms, separately for each stimulus type. The epochs were artefact rejected using the pop_autorej function from the EEGLAB toolbox using MATLAB software. EFRs were computed by adding responses to the positive and negative

stimulus polarities (Add = (Pos + Neg)/2). All subsequent correlational analyses were conducted on the averaged cABR waveforms.

The signal-to-noise ratio (SNR) of the cABR responses was calculated as 20log10(RMSpost/RMSpre), where RMSpost and RMSpre were the root mean square of the amplitude in the 'response' period (defined as 25-75 ms post-stimulus) and the root mean square of the pre-stimulus period (defined as 50 ms pre-stimus until stimulus onset at 0 ms) respectively. Three participants were removed from the EEG analysis as they had SNRs < 1.5 dB due to movement and other muscle artefact.

Stimulus-to-response cross-correlations (r-values) were generated using established MATLAB routines (Brainstem Toolbox 2013) (Skoe & Kraus, 9000). The maximum cross-correlation values were chosen irrespective of the lag. Cross-correlations were performed against the Hilbert envelope of the stimulus, as recorded through the transducer and an artificial ear (GRAS Type 43-AG). All data transformation and statistical tests (correlational analyses) were conducted using Minitab®.

3 Results

3.1 Hearing thresholds and age.

Figure 2 shows how hearing sensitivity generally declined slightly with age. The correlation between age and the pure-tone average hearing loss (at 500 Hz, 1 kHz, 2 kHz and 4kHz) was significant for the left ears, $r = .44$, $p = .008$, but not for the right ears.

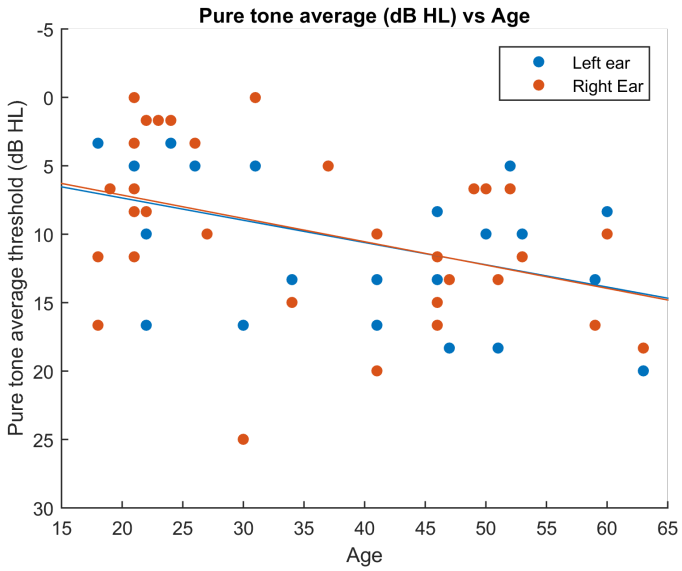


Figure 3. Pure tone average hearing thresholds for the left and right ears (average at 500, 1k, 2k, 4k Hz) plotted as a function of age. The blue and red lines indicate the least-squares regression line.

3.2 Correlations: QuickSIN and TFS with age.

Figure 3 shows the relationships between age, speech in noise perception, and TFS sensitivity. While speech in noise performance declined gradually with age (and was not significant at the .05 alpha level), TFS sensitivity declined rapidly and the relationship was highly significant ($r = -.57$, $p < .001$).

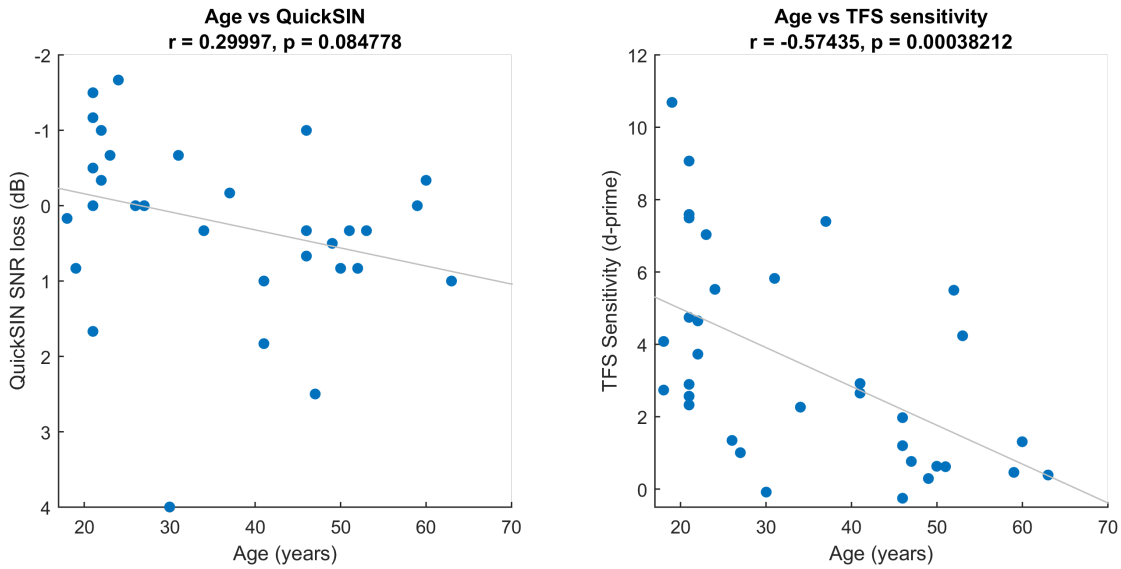


Figure 4. Left panel: Speech in noise performance (QuickSIN SNR loss) as a function of age. Note that the direction of the y-axis has been reversed so that better performance is up. The grey line indicates a least-squares linear regression. Right panel: Sensitivity to temporal fine structure (TFS) as a function of age. TFS sensitivity is expressed using a d' measure. The grey line indicates a least-squares linear regression

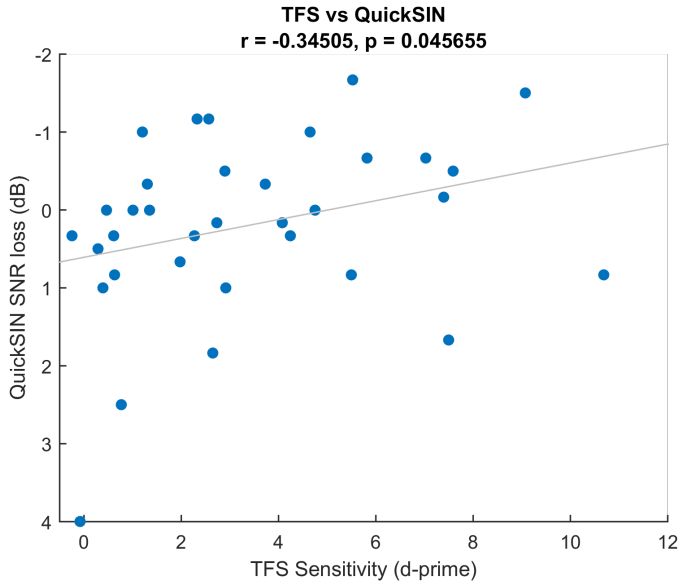


Figure 5. Speech in noise performance (QuickSIN) as a function of Temporal Fine Structure (TFS) sensitivity (expressed as d'). Note that the direction of the y-axis has been reversed so that better performance is up. The grey line indicates a least-squares linear regression.

3.3 Correlations: Speech in noise with TFS sensitivity

Figure 4 shows how speech in noise performance was moderately and significantly related to TFS sensitivity (r = -.34, p = .045). Participants who had good TFS sensitivity generally had good speech in noise scores.

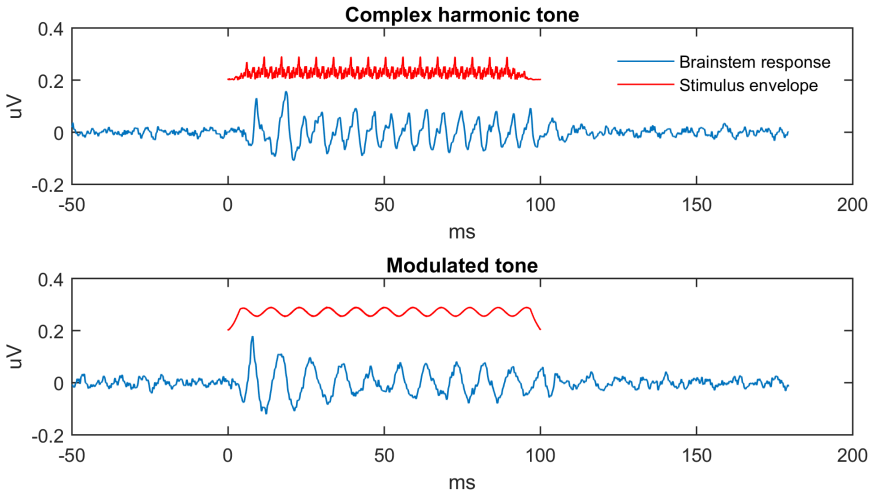


Figure 6. Grand average (across all participants) brainstem responses for the complex harmonic tone (top panel) and modulated tone (bottom panel). The Hilbert envelope of the stimulus waveform is shown in red.

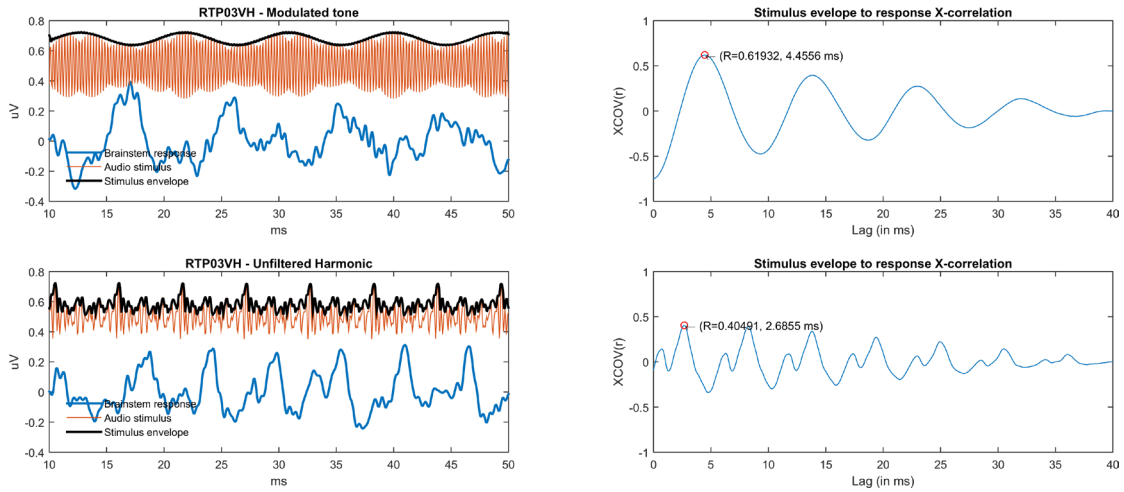


Figure 7. An illustration for a single listener of the calculation of the cross-correlation values for the modulated tone (top panels) and complex harmonic tone (bottom panels). The left panels show the brainstem responses (in blue) and the stimulus waveform (red) and its Hilbert envelope (black). Note the stimulus waveform here is not the electrical signal: rather it has been recorded through the transducer and an artificial ear. The right panels show the cross-correlation values between the stimulus envelope and the brainstem response as a function of the lag between them. The maximum of the cross-correlation function was determined for further analysis.

3.4 Electrophysiology

Grand average responses for the complex harmonic and modulated tones are shown in Figure 6. The responses show clear phase-locking to periodicity in the stimuli.

A measure of stimulus encoding strength was generated by calculating the cross-correlation between the stimulus envelope (as measured through the transducer and artificial ear) and the brainstem response. Figure 7 shows this process for one listener.

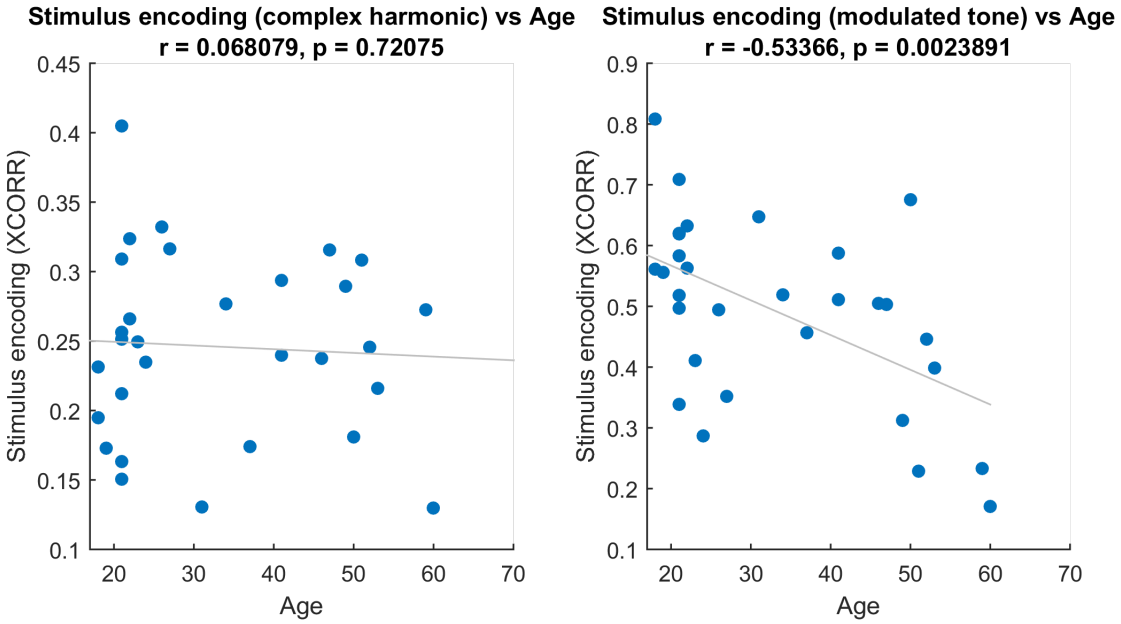


Figure 8. Stimulus-to-response cross-correlation values between the complex harmonic (left) and modulated (right) stimuli and brainstem response as a function of age.

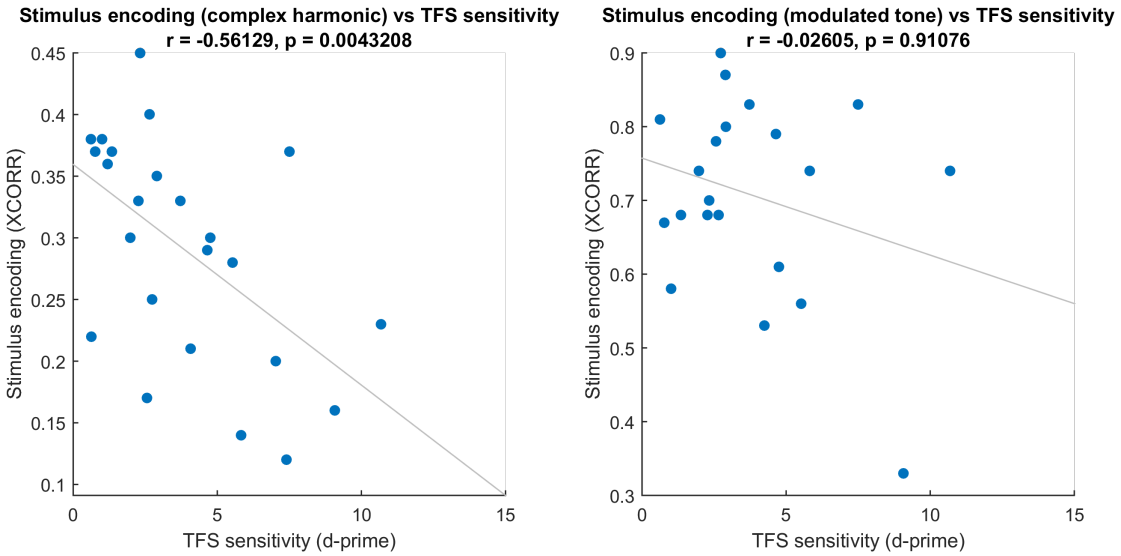


Figure 9. Stimulus-to-response cross-correlation values between the complex harmonic (left) and modulated (right) stimuli and brainstem response as a function of TFS sensitivity (expressed as d' scores).

3.5 Correlations: Stimulus encoding accuracy with age, TFS sensitivity and QuickSIN

The maximum cross-correlation value obtained from each participant and stimulus type was then correlated with age, the TFS sensitivity and speech in noise scores. Figure 8 shows that increasing age was associated with decreasing stimulus encoding accuracy, but only for the modulated tone stimulus (r = -.53, p < .001).

Figure 9 shows a striking relationship between increased TFS sensitivity and reduced stimulus encoding for the complex harmonic tone (r = -.56, p = .004). Interestingly, there was no such relationship for the modulated tone. There were also no significant relationships between stimulus encoding accuracy and the speech in noise scores (Figure 10).

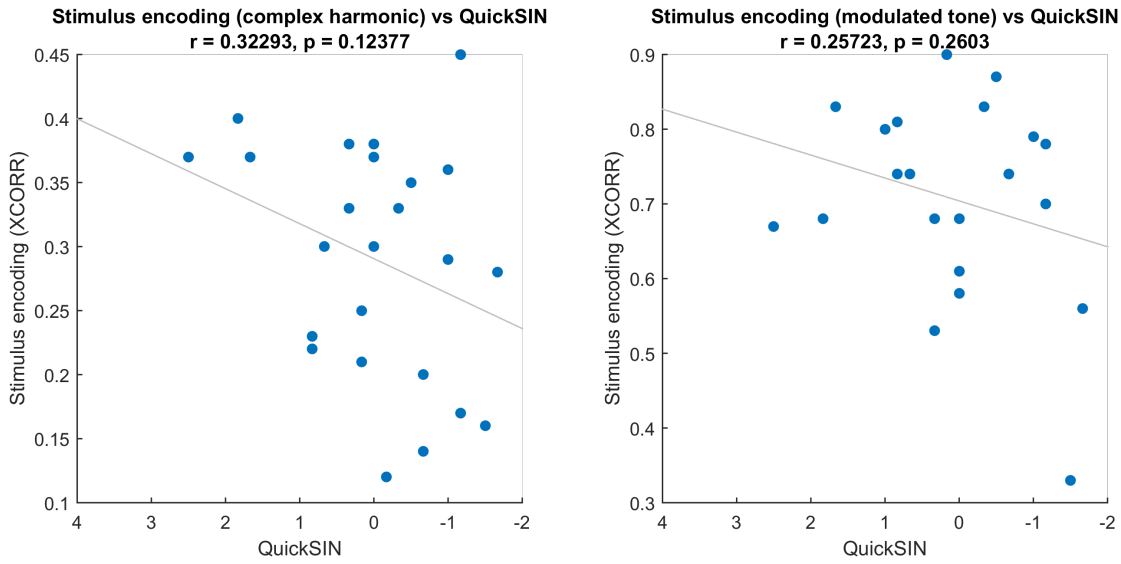


Figure 10. Stimulus-to-response cross-correlation values between the complex harmonic (left) and modulated (right) stimuli and brainstem response as a function of speech in noise scores.

3.6 Regression analysis

In order to determine which psychophysical and/or EEG measure best predicted the speech scores, best subsets regression was performed. The QuickSIN scores were entered as the response variable. Age, four-frequency pure-tone hearing thresholds (in the better ear), TFS sensitivity, and stimulus encoding accuracy for both the complex harmonic and modulated tones were entered as predictors. The regression indicated that QuickSIN scores were best predicted by a combination of all variables except the hearing thresholds (R^2 = 64.4, $F_{(4,26)}$ = 6.32, p = .004). A follow-up standard linear regression using the four predictors identified by the best-subsets procedure echoed these results, although TFS sensitivity was not a significant factor. The main model was significant (R^2 = 41.8, $F_{(4,26)}$ = 4.49, p = .007), with age accounting for most of the variance in the model (18.1%), followed by the stimulus encoding accuracy variables contributing 13.7% and 10.0% for the modulated and complex harmonic tones respectively. When the pure-tone average hearing threshold were added to the model, they contributed only 2% variance.

4 Discussion

In summary, we found that in our group of normally-hearing adults (with a variety of ages), speech-in-noise performance was negatively correlated with TFS sensitivity and age. TFS sensitivity was also positively correlated with stimulus encoding accuracy for the complex harmonic stimulus, while increasing age was associated with lower stimulus encoding accuracy for the modulated tone stimulus. Surprisingly, we found that better speech in noise understanding was associated with worse stimulus encoding accuracy. Despite this unexpected direction of correlation the measures did contribute modestly to a regression model predicting the speech in noise scores. A regression analysis found that age and the combination of the two stimulus encoding accuracy measures had roughly equal contributions to the model.

Further work in this area should consider other psychophysical predictors that are known to be associated with speech understanding, such as measures of temporal modulation sensitivity, and EEG measures that more closely match the stimuli used in the psychophysics.

Acknowledgments

Supported provided by the National Health and Medical Research Council of Australia (Peter Doherty Early Career Research Fellowship #1069999 to HIB). The Bionics Institute acknowledges the support it receives from the Victorian Government through its Operational Infrastructure Support Program. HIB is currently supported by a FWO Pegasus Marie Curie Fellowship at KU Leuven, Belgium.

References

Alaerts, J., Luts, H., Van Dun, B., Desloovere, C., & Wouters, J. (2010). Latencies of auditory steady-state responses recorded in early infancy. *Audiology and Neuro-Otology, 15*(2), 116-127.

Audiology, A. A. o. (2003). Pediatric Amplification Protocol. Washington (DC).

Billings, C. (2012). Clinical Use of Aided Cortical Auditory Evoked Potentials as a Measure of Physiological Detection or Physiological Discrimination. *International Journal of Otolaryngology, 2012*.

Billings, C. J., Tremblay, K. L., & Miller, C. W. (2011). Aided cortical auditory evoked potentials in response to changes in hearing aid gain. *International Journal of Audiology, 50*(7), 459-467. doi: 10.3109/14992027.2011.568011

Billings, C. J., Tremblay, K. L., Souza, P. E., & Binns, M. A. (2007). Effects of hearing aid amplification and stimulus intensity on cortical auditory evoked potentials. *Audiology and Neuro-Otology, 12*(4), 234-246. doi: 10.1159/000101331

Carter, L., Golding, M., Dillon, H., & Seymour, J. (2010). The detection of infant cortical auditory evoked potentials (CAEPs) using statistical and visual detection techniques. [Comparative Study

Research Support, Non-U.S. Gov't]. *Journal of the American Academy of Audiology, 21*(5), 347-356. doi: 10.3766/jaaa.21.5.6

Chang, H.-W., Dillon, H., Carter, L., Van Dun, B., & Young, S.-T. (2012). The relationship between cortical auditory evoked potential (CAEP) detection and estimated audibility in infants with sensorineural hearing loss *International Journal of Audiology, 51*, 663-670.

Hopkins, K., & Moore, B. C. (2007). Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information. *Journal of the Acoustical Society of America, 122*(2), 1055-1068. doi: 10.1121/1.2749457

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 116*(4 Pt 1), 2395-2405.

Luts, H., Desloovere, C., & Wouters, J. (2006). Clinical application of dichotic multiple-stimulus auditory steady-state responses in high-risk newborns and young children. *Audiology and Neuro-Otology, 11*(1), 24-37.

Luts, H., Desloovere, C., Kumar, A., Vandermeersch, E., & Wouters, J. (2004). Objective assessment of frequency-specific hearing thresholds in babies. *International Journal of Pediatric Otorhinolaryngology, 68*(7), 915-926.

Moore, B. C. J., & Sek, A. (2009). Development of a fast method for determining sensitivity to temporal fine structure. *International Journal of Audiology, 48*, 161-171.

Moore, B. C. J., & Sek, A. (2012). Implementation of two tests for measuring sensitivity to temporal fine structure *International Journal of Audiology, 51*, 58-63.

Skoe, E., & Kraus, N. (9000). Auditory Brain Stem Response to Complex Sounds: A Tutorial. *Ear and Hearing, Publish Ahead of Print*, 10.1097/AUD.1090b1013e3181cdb1272-1010.1097/AUD.1090b1013e3181cdb1272.

Chapter 34

Connectivity in Language Areas of the Brain in Cochlear Implant Users as Revealed by fNIRS

Colette M. McKay[1,2], Adnan Shah[1,3], Abd-Krim Seghouane[3], Zhou Xin[1,2], William Cross[1,4], Ruth Litovsky[5]

[1] *The Bionics Institute of Australia, Melbourne, Australia*
[2] *The University of Melbourne, Department of Medical Bionics, Melbourne, Australia*
[3] *The University of Melbourne, Department of Electrical and Electronic Engineering, Melbourne, Australia*
[4] *The University of Melbourne, Department of Medicine, Melbourne, Australia*
[5] *The University of Wisconsin-Madison, Waisman Center, Madison, USA*

Contact author: CMM, cmckay@bionicsinstitute.org

Abstract

Many studies, using a variety of imaging techniques, have shown that deafness induces functional plasticity in the brain of adults with late-onset deafness, and in children changes the way the auditory brain develops. Cross modal plasticity refers to evidence that stimuli of one modality (e.g. vision) activate neural regions devoted to a different modality (e.g. hearing) that are not normally activated by those stimuli. Other studies have shown that multimodal brain networks (such as those involved in language comprehension, and the default mode network) are altered by deafness, as evidenced by changes in patterns of activation or connectivity within the networks. In this paper, we summarise what is already known about brain plasticity due to deafness and propose that functional near-infra-red spectroscopy (fNIRS) is an imaging method that has potential to provide prognostic and diagnostic information for cochlear implant users. Currently, patient history factors account for only 10% of the variation in post-implantation speech understanding, and very few post-implantation behavioural measures of hearing ability correlate with speech understanding. As a non-invasive, inexpensive and user-friendly imaging method, fNIRS provides an opportunity to study both pre- and post-implantation brain function. Here, we explain the principle of fNIRS measurements and illustrate its use in studying brain network connectivity and function with example data.

Keywords: fNIRS, cochlear implants, deafness, brain plasticity, connectivity in brain networks

1 Introduction

1.1 Deafness, Language and Brain Plasticity: Evidence from Imaging Studies

Speech understanding involves complex multimodal networks involving vision, hearing and sensory motor areas as well as memory and frontal lobe functions mostly in the left hemisphere (LH), and encompasses a range of elements such as phonology, semantics, and syntactics. The right hemisphere (RH) has fewer specialised functions for language processing, and its role seems to be mostly in the evaluation of the communication context (Vigneau et al., 2011).

Multiple imaging studies have shown that adults who have undergone periods of profound post-lingual deafness demonstrate changes in brain activity and function in language-associated brain areas that are not observed in normally-hearing individuals, and that further functional plasticity occurs as a result of cochlear implantation.

Lee et al. (2003) used positron emission tomography (PET) to compare resting-state activity in 9 profoundly deaf individuals and 9 age-matched normal-hearing controls. They found that glucose metabolism in some auditory areas (anterior cingulate gyri - Brodmann area 24 [BA24] - and superior temporal cortices - BA41, BA42) was lower than in normally-hearing people, but significantly increased with duration of deafness, and concluded that plasticity occurs in the sensory-deprived mature brain. In a later paper the same group showed, using PET resting state data, that children with good speech understanding 3 years after implantation had enhanced metabolic activity in the left prefrontal cortex and decreased metabolic activity in right Heschl's gyrus and in the posterior superior temporal sulcus before implantation compared to those with poor speech understanding (Lee et al., 2007). They argued that increased activity in the resting state in auditory areas may reflect cross modal plasticity that is detrimental to later success with the CI. A recent paper by Dewey and Hartley (2015) using functional near infrared spectroscopy (fNIRS) has supported the proposal that cross modal plasticity occurs in deafness: they demonstrated that the auditory cortex of deaf individuals is abnormally activated by simple visual stimuli (moving or flashing dots).

Not all studies have suggested detrimental effects of brain changes due to deafness on the ability to adapt to listening with a CI. Giraud et al used PET to study the activity induced by speech stimuli in CI users, and found that, compared to normal-hearing listeners, they had altered functional specificity of the superior temporal cortex, and exhibited contribution of visual regions to sound recognition (Giraud, Price, Graham, & Frackowiak, 2001). Furthermore, the contribution of the visual cortex to speech recognition *increased* over time after implantation (Giraud, Price, Graham, Truy, et al., 2001), suggesting that the CI users were actively using enhanced audio-visual integration to facilitate their learning of the novel speech sounds received through the CI. In contrast, Rouger et al (2012) suggested a negative impact of cross modal plasticity: they found that the right temporal voice area (TVA) was abnormally activated in CI users by a *visual* speech-reading task and that this activity declined over time after implantation while the activity in Broca's area (normally activated by speech reading) increased over time after implantation. Coez et al (2008) also used PET to study activation by *voice* stimuli of the TVA in CI users with poor and good speech understanding. The voice stimuli induced bilateral activation of the TVA along the superior temporal sulcus in both normal hearing listeners and CI users with good speech understanding, but not in CI users with poor understanding. This result is consistent with the proposal of Rouger that the TVA is 'taken over' by visual speech reading tasks in CI users who do not understand speech well. Strelnikov et al. (2013) measured PET resting state activity and activations elicited by auditory and audio-visual speech in CI users soon after implantation and found that good speech understanding after 6 months of implant use was predicted by a higher activation level of the right occipital cortex and a lower activation in the right middle superior temporal gyrus. They suggested that the pre-implantation functional changes due to reliance on lip-reading were *advantageous* to development of good speech understanding through a CI via enhanced audio-visual integration. The overall picture that comes from these studies, suggests that the functional changes that occur during deafness can both enhance and degrade the ability to adapt to CI listening, and this may depend on the strategies the each person uses while deaf to facilitate their communication.

Lazard et al. have published a series of studies using functional magnetic resonance imaging (fMRI), in which they related pre-implant data with post-implant speech understanding. They found that, when doing a rhyming task with written words, CI users with later good speech understanding showed an activation pattern that was consistent with them using the normal phonological pathway to do the task, whereas those with poor outcomes used a route normally associated with lexical-semantic understanding (Lazard et al., 2010). They further compared

activity evoked by speech and non-speech imageries in the right and left posterior superior temporal gyrus/supramarginal gyrus (PSTG/SMG) (Lazard et al., 2013). These areas are normally specialised for phonological processing in the left hemisphere and environmental sound processing in the right hemisphere. They found that responses in the left PSTG/SMG to speech imagery and responses in the right PSTG/SMG to environmental sound imagery both declined with duration of deafness, as memories of these sounds are lost. However, abnormally high activity was observed in response to phonological visual items in the *right* PSTG/SMG (contralateral to the zone where phonological activity decreased over time). They suggested that this pattern may be explained by the recruiting of the right PSTG/SMG region for phonological processing, rather than it being specialised for processing of environmental sounds.

In summary, the studies reviewed above have shown functional changes due to periods of deafness, some of which are detrimental to post-implant speech understanding. On the other hand increased activity in the occipital lobe in resting state and for auditory or audio-visual stimuli was positively correlated with post-implant speech understanding. It is evident that the functional changes noted by the authors involve not just the auditory cortex, but the distributed multimodal language networks and that the reliance on lip-reading while deaf may be a major factor that drives the functional changes.

1.2 Functional Near Infra-red Spectroscopy

Light in the near infra-red spectrum (wavelengths 650-1000 nm) is relatively transparent to human tissues. The absorption spectra of oxygenated and de-oxygenated haemoglobin (HbO and HbR respectively) for near infra-red light differ in that HbO maximally absorbs light of longer wavelength (900-1000 nm) whereas HbR maximally absorbs light of shorter wavelength. These differential absorption spectra make it possible, when using near infra-red light of different wavelengths to image cortical activity, to separate out changes in the concentration of HbO and HbR. In response to stimulus-driven activity in the brain, an increase in oxygenated blood is directed to the region. When the supply of oxygenated blood exceeds the demand for oxygen, the concentration of HbR decreases. Thus a pattern emerges of HbO and HbR concentrations changing in opposite directions in response to neural activity. The measurement of changes in HbO and HbR resulting from a neural response to a stimulus or due to resting state activity can be analysed to produce activation patterns and also to derive connectivity measures between intra-hemispheric and cross-hemispheric regions of interest.

In the fNIRS imaging system optodes are placed at various locations on the scalp. Multiple channels can be used, and each fNIRS channel consists of a two optodes: a source and a detector. The source emits a light beam into the brain, directed perpendicular to the scalp surface, and the detector detects the light emerging from the brain. The detected light beam has scattered in a banana-shaped pathway through the brain reaching a depth of approximately half the distance between the source and detector (see Figure 1). The distance between source and detector, therefore, will determine the depth of cortical activity that is measured, and the amount of light that can be detected from the source at the detector position. It is currently thought that an optimal compromise between imaging depth and strength of the signal at the detector is obtained with a 3-cm distance of source and detector. In a montage of multiple optodes, each source can be associated with a number of surrounding detectors to form a multi-channel measurement system. In continuous wave systems, two frequencies of light are emitted by the diodes, and the signals at each source are frequency-modulated at different rates to facilitate separation of the light from different sources at the same detector position.

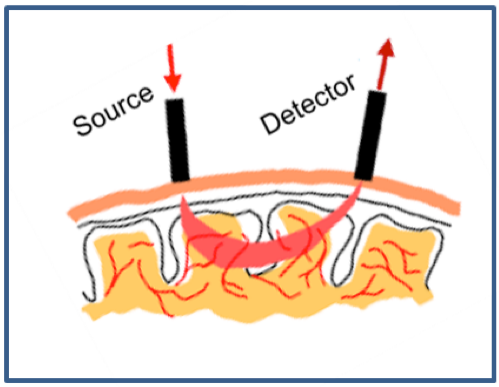


Figure 1. Schematic of fNRIS optodes and light path in a single channel

For studying language areas in the brain in CI users, fNIRS offers some advantages over other imaging methods: compared to PET it is non-invasive; in contrast to fMRI it can be used easily with implanted devices, is silent, and is more robust to head movements; compared to EEG/MEG it has much greater spatial resolution, and in CI users is free from electrical or magnetic artifacts from the device or the stimuli. The portability, low-cost, and patient-friendly nature of fNIRS (similar to EEG) makes it a plausible method to contribute to routine clinical management of patients, including infants and children.

However, there are also limitations of fNIRS, which means that the research or clinical questions that can be validly investigated with it need to be carefully considered. The spatial resolution is limited by the density of the optodes used, and is not generally as good as that found with fMRI or PET imaging. Importantly, the depth of imaging is limited, so that it is only suitable for imaging areas of the cortex near the surface [although other designs of fNIRS systems that use lasers instead of diodes and use pulsatile stimuli can derive 3-dimension images of the brain, at least in infants (Cooper et al., 2014)]. However, fNIRS has been successfully used in a range of studies to determine effects on language processing in hearing populations of adults and children (Quaresima et al., 2012), therefore it shows great promise for assessing language processing in deaf populations and those with cochlear implants. The ability of fNIRS to image language processing in CI users has been demonstrated (Sevy et al., 2010).

In this paper we describe methods and present preliminary data for two investigations using fNIRS to compare CI users and normally-hearing listeners. In the first experiment, we measure resting state connectivity, and in the second experiment we compare the activation of cortical language pathways by visual and auditory speech stimuli and their connectivity.

2 Methods

2.1 fNIRS equipment and data acquisition

Data were acquired using a multichannel 32 optode (16 sources and 16 detectors) NIRScout system. Each source LED emitted light of two wavelengths (760 and 850 nm). The sources and detectors were mounted in a pre-selected montage using an EASYCAP with grommets to hold the fNIRS optodes, which allowed registration of channel positions on a standard brain template using the international 10-20 system. To ensure optimal signal detection in each channel, the cap was fitted first, and the hair under each grommet moved aside before placing the optodes in the grommets. For CI users, the cap was fitted over the transmission coil with the speech processor hanging below the cap. The data were exported into MatLab or NIRS-SPM for analysis.

2.2 Resting-State Connectivity in CI users using fNIRS

In this study we are comparing the resting state connectivity of a group of normal hearing listeners compared to a group of experienced adult CI users. Brain function in the resting state, or its default-mode activity (Raichle & Snyder, 2007) is thought to reflect the ability of the brain to predict changes to the environment and track any deviation from the predicted. Resting state connectivity (as measured by the correlation of resting state activity between different cortical regions) is influenced by functional organisation of the brain, and hence is expected to reflect plastic changes such as those due to deafness. In this study, we hypothesised that CI users would exhibit resting-state connectivity that was different from that of normally-hearing listeners.

We used a 4x4 montage of 8 sources and detectors (24 channels) in each hemisphere (Figure 2), which covered the auditory and somatosensory regions of the brain. Sources and detectors were separated by 3 cm. Data in each channel were pre-processed to remove 'glitches' due to head movements and 'good' channels were identified by a significant cross-correlation ($r > 0.75$) between the data for the two wavelengths. Any 'poor' channels were discarded. The data were then converted to HbO and HbR based on a modified Beer-Lambardt Law (Cope et al., 1988).

Since resting state connectivity is based upon correlations between activity in relevant channels, it is important to carefully remove any aspects of the data (such as fluctuations from heartbeat, breathing, or from movement artifacts that are present in all channels) that would produce a correlation but is not related to neural activity. These processing steps to remove unwanted signals from the data included: minimizing regional drift using a discrete cosine transform; removing global drift using PCA 'denoising' techniques, low-pass filtering (0.08 Hz) to remove heart and breathing fluctuations. Finally the connectivity between channels was calculated using Pearson's correlation, r. To account for missing channels, the data were reduced to 9 regions of interest in each hemisphere by averaging the data from groups of 4 channels.

2.3 Language Networks in CI users using fNIRS

In the second study, we aimed to measure the activity in, and connectivity between, regions in the brain involved in language processing in response to auditory and visual speech stimuli. We hypothesised that, compared to normal-hearing listeners, CI users would show altered functional organisation, and furthermore that this difference would be correlated with lip-reading ability and with their auditory speech understanding.

The optode montage for this experiment is shown in Figure 3. fNIRS data were collected using a block design in two sessions. In the first session, spondee words were used in 12-s blocks of visual spondees alternating with blocks of auditory spondees, with blocks separated by 15-25 s silent gaps. Between the stimulus blocks participants were asked to focus on a black screen in front of them, which was also used to present the visual stimuli. The auditory stimuli were presented via EAR-4 insert ear phones in normally-hearing listeners, or via direct audio input for the CI users. Sounds were only presented to the right ear of subjects (the CI users were selected to have unilateral right implants). In the second session, blocks consisted of sentences alternating between auditory-alone and audio-visual.

3 Results

3.1 Preliminary fNIRS Data for Resting-State Connectivity

Figure 2 shows example data for resting state connectivity in one CI user and one normal-hearing (NH) listener. The NH data is representative of all NH listeners tested to date, and shows particularly strong connectivity between analogous regions in left and right hemispheres: a result consistent with other reports from studies investigating resting state connectivity using fNIRS (Medvedev, 2014). To date, all the CI users, like the example shown, have consistently

showed lower inter-hemispheric connectivity than the NH listeners. Further work is being conducted to evaluate group differences in more detail.

3.2 Preliminary fNIRS Data for Language Networks

Figure 3 shows example data from one NH participant in the second experiment showing activation patterns and haemodynamic response function in response to visual spondees. The activation pattern shows expected bilateral activation of Broca's area (which is associated with articulation of speech), and some activation of language-related areas in the left hemisphere. Further work is being undertaken to compare activation patterns and connectivity in language pathways between CI users and NH listeners, and to correlate relevant differences to individual behavioural measures of speech understanding and lip-reading ability.

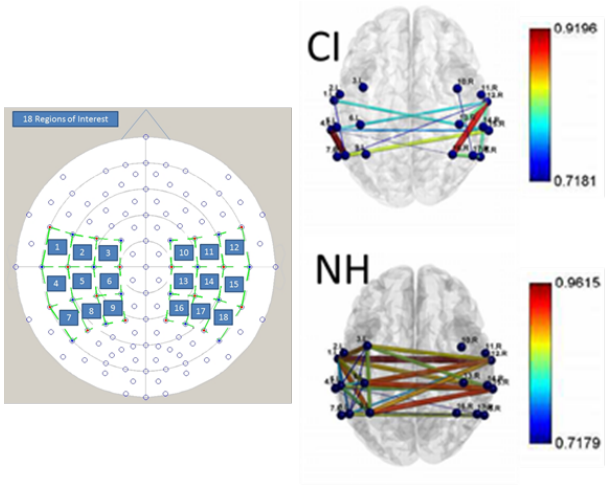


Figure 2. Example data showing resting state connectivity. The left panel shows the montage used and the 18 regions of interest. The right panels show the connectivity maps: the line colour and thickness denote the strength of connectivity (r) in the different brain regions of interest (colour bar denotes r-values). The data clearly show a lack of inter-hemispheric connectivity in the CI user compared to the NH listener.

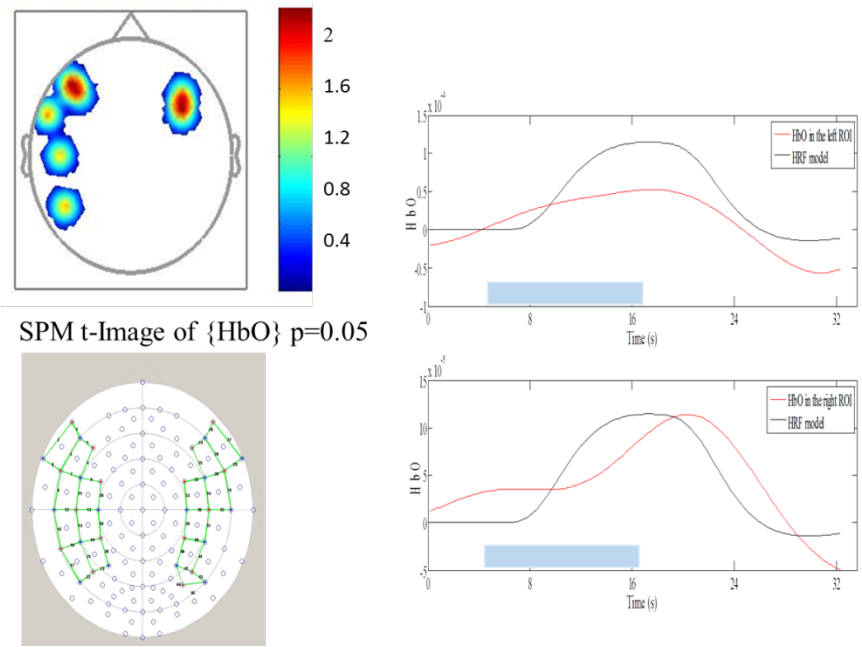


Figure 3. Left panels show the montage of optodes used in Experiment 2 and the activation pattern elicited in one NH listener when watching visual spondees. The colour bar shows the t-values. The right panels show the average bilateral haemodynamic response functions from the regions of greatest activation in each hemisphere. The blue blocks indicate 'stimulus-on' times.

4 Discussion

Our preliminary data shows that fNIRS has the potential to provide insight into the way that brain functional organisation is altered by deafness and subsequent cochlear implantation. The fNIRS tool may be particularly useful for longitudinal studies that track changes over time, so that the changes in brain function can be correlated with simultaneous changes in behavioural performance. In this way, new knowledge about functional organisation and how it relates to individual's ability to process language can be gained. For example, an fNIRS pre-implant test may, in the future, provide valuable prognostic information for clinicians and patients, and provide guidance for individual post-implant therapies designed to optimise outcomes. Routine clinical use of fNIRS is feasible due to its low-cost. It will be particularly useful in studying language development in deaf children due to its patient-friendly nature.

Acknowledgments

This research was supported by a **veski** fellowship to CMM, an Australian Research Council Grant (FT130101394) to AKS, a Melbourne University PhD scholarship to ZX, the Australian Fulbright Commission for a fellowship to RL, the Lions Foundation, and the Melbourne Neuroscience Institute. The Bionics Institute acknowledges the support it receives from the Victorian Government through its Operational Infrastructure Support Program.

References

Coez, A., Zilbovicius, M., Ferrary, E., Bouccara, D., Mosnier, I., Ambert-Dahan, E., Bizaguet, E., Syrota, A., Samson, Y., & Sterkers, O. (2008). Cochlear implant benefits in deafness rehabilitation: PET study of temporal voice activations. *Journal of nuclear medicine, 49*(1), 60-67. doi: 10.2967/jnumed.107.044545

Cooper, R. J., Magee, E., Everdell, N., Magazov, S., Varela, M., Airantzis, D., Gibson, A. P., & Hebden, J. C. (2014). MONSTIR II: a 32-channel, multispectral, time-resolved optical tomography system for neonatal brain imaging. *The Review of scientific instruments, 85*(5), 053105. doi: 10.1063/1.4875593

Cope, M., Delpy, D. T., Reynolds, E. O., Wray, S., Wyatt, J., & van der Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data. *Advances in experimental medicine and biology, 222*, 183-189.

Dewey, R. S., & Hartley, D. E. (2015). Cortical cross-modal plasticity following deafness measured using functional near-infrared spectroscopy. *Hearing Research*. doi: 10.1016/j.heares.2015.03.007

Giraud, A. L., Price, C. J., Graham, J. M., & Frackowiak, R. S. (2001). Functional plasticity of language-related brain areas after cochlear implantation. *Brain, 124*(Pt 7), 1307-1316.

Giraud, A. L., Price, C. J., Graham, J. M., Truy, E., & Frackowiak, R. S. (2001). Cross-modal plasticity underpins language recovery after cochlear implantation. *Neuron, 30*(3), 657-663.

Lazard, D. S., Lee, H. J., Truy, E., & Giraud, A. L. (2013). Bilateral reorganization of posterior temporal cortices in post-lingual deafness and its relation to cochlear implant outcome. *Human Brain Mapping, 34*(5), 1208-1219. doi: 10.1002/hbm.21504

Lazard, D. S., Lee, H. J., Gaebler, M., Kell, C. A., Truy, E., & Giraud, A. L. (2010). Phonological processing in post-lingual deafness and cochlear implant outcome. *Neuroimage, 49*(4), 3443-3451. doi: 10.1016/j.neuroimage.2009.11.013

Lee, H. J., Giraud, A. L., Kang, E., Oh, S. H., Kang, H., Kim, C. S., & Lee, D. S. (2007). Cortical activity at rest predicts cochlear implantation outcome. *Cereb Cortex, 17*(4), 909-917.

Lee, J. S., Lee, D. S., Oh, S. H., Kim, C. S., Kim, J. W., Hwang, C. H., Koo, J., Kang, E., Chung, J. K., & Lee, M. C. (2003). PET evidence of neuroplasticity in adult auditory cortex of postlingual deafness. *Journal of Nuclear Medicine, 44*(9), 1435-1439.

Medvedev, A. V. (2014). Does the resting state connectivity have hemispheric asymmetry? A near-infrared spectroscopy study. *Neuroimage, 85 Pt 1*, 400-407. doi: 10.1016/j.neuroimage.2013.05.092

Quaresima, V., Bisconti, S., & Ferrari, M. (2012). A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults. *Brain and language, 121*(2), 79-89. doi: 10.1016/j.bandl.2011.03.009

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea. *Neuroimage, 37*(4), 1083-1090; discussion 1097-1089. doi: 10.1016/j.neuroimage.2007.02.041

Rouger, J., Lagleyre, S., Demonet, J. F., Fraysse, B., Deguine, O., & Barone, P. (2012). Evolution of crossmodal reorganization of the voice area in cochlear-implanted deaf patients. *Human Brain Mapping, 33*(8), 1929-1940. doi: 10.1002/hbm.21331

Sevy, A. B., Bortfeld, H., Huppert, T. J., Beauchamp, M. S., Tonini, R. E., & Oghalai, J. S. (2010). Neuroimaging with near-infrared spectroscopy demonstrates speech-evoked activity in the auditory cortex of deaf children following cochlear implantation. *Hearing Research, 270*(1-2), 39-47.

Strelnikov, K., Rouger, J., Demonet, J. F., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2013). Visual activity predicts auditory recovery from deafness after adult cochlear implantation. *Brain, 136*(Pt 12), 3682-3695. doi: 10.1093/brain/awt274

Vigneau, M., Beaucousin, V., Herve, P. Y., Jobard, G., Petit, L., Crivello, F., Mellet, E., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. *Neuroimage, 54*(1), 577-593. doi: 10.1016/j.neuroimage.2010.07.036

Chapter 35

Isolating Neural Indices of Continuous Speech Processing at the Phonetic Level

Giovanni M. Di Liberto[1,2,3], Edmund C. Lalor[1,2,3]

[1] *Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland*
[2] *School of Engineering, Trinity College Dublin, Dublin, Ireland*
[3] *Trinity Centre for Bioengineering, Trinity College Dublin, Dublin, Ireland*

Contact author: G.D.L., diliberg@tcd.ie

Abstract

The human ability to understand speech across an enormous range of listening conditions is underpinned by a hierarchical auditory processing system whose successive stages process increasingly complex attributes of the acoustic input. In order to produce a categorical perception of words and phonemes, it has been suggested that, while earlier areas of the auditory system undoubtedly respond to acoustic differences in speech tokens, later areas must exhibit consistent neural responses to those tokens. Neural indices of such hierarchical processing in the context of continuous speech have been identified using low-frequency scalp-recorded electroencephalography (EEG) data. The relationship between continuous speech and its associated neural responses has been shown to be best described when that speech is represented using both its low-level spectrotemporal information and also the categorical labelling of its phonetic features (Di Liberto et al., submitted). While the phonetic features have been proven to carry extra-information not captured by the speech spectrotemporal representation, the causes of this EEG activity remain unclear. This study aims to demonstrate a framework for examining speech-specific processing and for disentangling high-level neural activity related to intelligibility from low-level activity in response to spectrotemporal fluctuations of speech. Preliminary results suggest that neural measure of processing at the phonetic level can be isolated.

Keywords: hierarchical, intelligibility, EEG, noise vocoding, priming, natural speech, phoneme

1 Introduction

Speech processing is an active cognitive activity underpinned by a complex hierarchical system. Recent studies with human neuroimaging have provided new important insights into speech and language processing which support this statement (Chang et al., 2010; Okada et al., 2010; Peelle et al., 2010; DeWitt & Rauschecker, 2012; Hickok, 2012). In particular, evidence for hierarchical speech processing emerged from functional magnetic resonance imaging (fMRI) and electrocorticography (ECoG) studies. Although these methodologies have provided important scientific insights, they have many limitations. For example, fMRI does not allow for the study of the fast temporal dynamics typical of continuous speech, and ECoG studies are limited to patients suffering from severe cases of epilepsy.

Electro- and magnetoencephalography (EEG/MEG), as macroscopic non-invasive technologies, have the potential for further progress on this topic. Traditionally, the low signal-to-noise ratio (SNR) has hampered the experiments conducted with these instruments, limiting the study of speech processing to repeated presentations of simple short sounds (e.g. discrete presentation of syllables), which elicit neural responses known as event related potentials (ERPs). Although studies based on ERPs have provided important insights on this topic, the human auditory processing mechanisms are tuned to process continuous speech and respond differently to discrete sounds (e.g. syllables in isolation, (Bonte et al., 2006)). Recent studies have provided encouraging evidence on the capability of EEG and MEG to track neural correlates of the low-frequency amplitude envelope of natural continuous speech (Aiken & Picton, 2008; Lalor & Foxe, 2010; Ding & Simon, 2012), a finding that has proven useful for investigating the mechanisms underlying speech processing (Luo & Poeppel, 2007) and its dependence on attention and multisensory integration (Kerlin et al., 2010; Ding & Simon, 2012; Power et al., 2012; Golumbic et al., 2013). However, the causes of this envelope tracking phenomenon remain unclear.

Our recent research effort has focused on the identification of neural indices of continuous speech processing at different levels of this hierarchical system. In particular, phonetic features of speech have been shown to capture information that does not emerge when using models based on the envelope, suggesting that EEG is indexing higher-level speech processing (Di Liberto et al., submitted). Here we seek to build on this work by disentangling the higher-level phonemic processing contributions from those related to the low-level acoustic properties of the speech signal. By utilising priming, we can control how much a degraded speech stimulus is recognised without changing the stimulus itself, exploiting the fact that speech recognition is an active process (Hickok, 2012). In doing so, we demonstrate a framework for examining speech-specific processing using EEG.

2 Methods

2.1 Subjects and Data Acquisition

Ten healthy subjects (5 male and 5 female, aged between 21 and 31) participated in this study. Electroencephalographic (EEG) data were recorded from 128 electrode positions, digitised at 512Hz using a BioSemi Active Two system. Monophonic audio stimuli were presented at a sampling rate of 44,100 Hz using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (*http://www.neurobs.com*). Testing was carried out in a dark room and subjects were instructed to maintain visual fixation on a crosshair centred on the screen, and to minimize eye blinking and all other motor activities for the duration of each trial.

2.2 Stimuli and Experimental Procedure

Audio-book versions of two classic works of fiction read by the same American English speaker were partitioned into speech snippets, each with a duration of 10 seconds. 120 snippets were randomly selected for the experiment. In order to alter the intelligibility of the speech, a method known as vocoding was implemented (Shannon et al., 1995). This method filters the speech into a number of frequency bands, and uses the amplitude envelope of each band to modulate band-limited noise. In our experiment, frequency bands were logarithmically-spaced (Greenwood, 1961) between 70 and 5000Hz. To determine the number of frequency bands to use in the vocoding procedure, an offline intelligibility test was performed for each subject using a separate set of 40 speech snippets. The number of frequency bands was chosen such that subjects could understand an average of two words in each 10s snippet.

The EEG experiments consisted of two conditions, standard trials and deviant trials, each containing the presentation of 3 speech snippets (Fig. 1). A **standard** trial consisted of a first snippet (snippet **a**) in which the presented speech was degraded using noise-vocoding with the

previously determined threshold. Therefore it was largely unintelligible. The second snippet (snippet **b**) contained the same speech, but in its original clear form. The noise-vocoded version was then played again in the third snippet (snippet **c**). As such, the third condition was affected by priming because the original clear speech was played immediately before and because the speech was vocoded at the limit of intelligibility. The speech snippets used to determine the number of vocoding frequency bands served also as a training stage in order to reduce the adaptation to the noise-vocoded speech during the experiment (Davis et al., 2005).

In order to measure the effect of priming, we also introduced deviant trials. These trials consisted of a modified version of **(a)** and/or **(c)**, where a random chunk of 5 seconds was replaced with words from a different trial. In both cases, the deviant probabilities were set to 10%. The participants were asked to identify the first and the second speech vocoded streams as a standard or a deviant presentation (after the presentations **(b)** and **(c),** respectively), using a level of confidence from 1 to 5 (from 'definitely a deviant' to 'definitely a standard').

2.3 EEG Data Analysis

The EEG data were inspected to identify channels with excessive noise or motor artefacts. Data were then digitally filtered using a Chebyshev type-2 band-pass filter with pass-band between 1 and 15 Hz and down-sampled to 128Hz. Also, data was referenced to the average of the two mastoid channels.

Linear regression (http://sourceforge.net/projects/aespa) was used to create a mapping between the EEG and five different representations of the speech stimuli:

1. The broadband amplitude **envelope (*Env*)** of the speech signal, which was calculated as $Env = |x_a(t)|, x_a(t) = x(t) + jx(t),$ where $x_a(t)$ is the complex analytic signal obtained by the sum of the original speech $x(t)$ and its Hilbert transform $x(t)$.
2. The **spectrogram (*Sgram*)** was obtained by partitioning the speech signal into three frequency bands logarithmically spaced between 70 and 5000 Hz according to Greenwood's equation (70 - 494 - 1680 – 5000 Hz, the same used for the vocoder), and computing the amplitude envelope for each band.
3. The **phonemic (*Ph*)** representation was computed using forced alignment (Yuan & Liberman, 2008), given a speech file and the correspondent textual orthographic transcription broken into 26 phonemes in the American English *International Phonetic Alphabet* (IPA). A multivariate time-series composed of 26 indicator variables was then obtained.
4. The **phonetic features (*Fea*)** encoding is a linear mapping of the phonetic representation into a space of 18 features (Mesgarani et al., 2014), which are a distinctive subset of the ones defined by Chomsky and Halle to describe the articulatory and acoustic properties of the speech phonetic content (Chomsky & Halle, 1968). In particular, the chosen features are related to the manner of articulation, to the voicing of a consonant, to the backness of a vowel, and to the place of articulation.
5. Finally, we propose a model that combines *Fea* and *Sgram* (***FS***) by applying a concatenation of the two representations.

For each representation of the speech, the result is a set of regression weights referred to as multivariate temporal response functions (TRFs). *k*-fold cross-validation ($k = 10$) was employed in order to build predictions of the EEG signal from the TRFs of each distinct speech representation model. The prediction accuracies were obtained as an average of the correlation values over a set of 12 best predictive electrodes (6 on the left side of the scalp and the symmetrical counterparts on the right), on a time-lag window from 0 to 250ms (Di Liberto et al., submitted).

The study was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

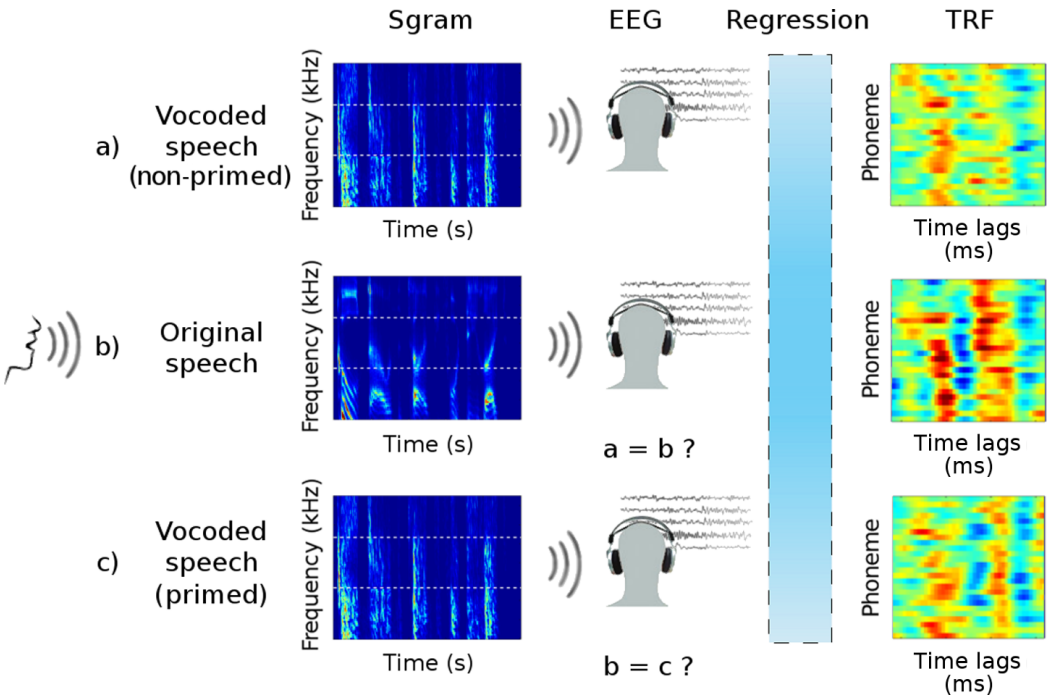


Figure 1. EEG data were recorded while subjects listened to groups of three 10 seconds long speech snippets. In the standard trial, the first (**a**) and third (**c**) speech streams of each group were the noise-vocoded version of the original one (**b**). The deviant presentations consisted of a modified version of (**a**) and/or (**c**). In both cases, the deviant probabilities were set to 10%. After the presentations (**b**) and (**c**), the participants were asked to identify the first and the second speech vocoded streams as a standard or a deviant presentation respectively. Linear regression was used to fit multivariate temporal response functions (TRFs) between the low-frequency EEG and five different representations of the speech stimulus.

3 Results

The main idea underlying the experimental paradigm was to take advantage of the priming effect to create two conditions in which the same speech stimuli were unintelligible and intelligible (condition **a** and **c** respectively; the prime is referred to as condition **b**). Such an experiment has the potential to disentangle acoustic and phonetic models and to allow us to address the non-overlapping cortical activity encoded by their TRFs.

3.1 Behavioural Results

The behavioural results (**Fig. 2A**) confirm that the experiment is creating these two conditions. Specifically, standard trials are detected with a higher confidence level (paired *t*-test, $p = 3.1 \times 10^{-5}$) while no difference emerged for deviant trials ($p = 0.41$). The distribution of responses (**Fig. 2B, top**) confirms that the participants were more confident in the detection of standard trials, therefore they could better understand the vocoded speech in the primed condition. Interestingly, the deviant trials were not confidently detected (**Fig. 2B, bottom**), in fact they consisted of a vocoded speech snippet partially inconsistent with the original one (~5 seconds). Therefore, the prime has no effect in the inconsistent portions and the subjects could not understand them better than they would have in the non-primed condition.

The hypothesis that subjects understand more the noise-vocoded speech if presented after the prime is met for standard trials, therefore the analysis of the recorded EEG data is conducted on them.

3.2 EEG Predictions and Model Comparison

128-channel EEG data were recorded from 10 subjects as they performed the task (**Fig. 1**). In this experiment, the original speech (condition **b**) was used as a prime for the second presentation of the noise-vocoded speech (condition **c**). Also, the results obtained in condition **b** provide a baseline with which to compare with previous research ((Di Liberto et al., submitted), in which EEG is used to obtain TRFs and prediction correlations from the individual phonemes in clear natural speech). Indeed, the main interest relies on the comparison between vocoded speech in the primed (**a**) and in the non-primed (**c**) conditions. In particular, we studied the standard trials, in which the difference between the primed and non-primed conditions can originate only in the EEG signal and it is related to an improved intelligibility (**Fig. 2A**).

TRF models were fit using linear regression on the standard for each speech representation (*Env*, *Sgram*, *Ph*, *Fea*, and *FS*) in the primed and non-primed condition. These models were then compared in terms of their ability of predicting the EEG data (see **Methods**). Interestingly, the phonetic models (*Ph* and *Fea*) show an improvement in their EEG predictions unseen for the ones based only on acoustical information (*Env* and *Sgram*, **Fig. 3**). This effect is driven by the whole distribution and not only by single subjects (Jack-knife analysis on the 10 subjects, two-way repeated measures ANOVA, Greenhouse-Geisser sphericity correction, df=1.8, F=66.6, $p < 0.0005$).

Additionally, given 50 seconds of EEG data belonging to the same condition, the TRFs have been employed to decode if they corresponded to primed or non-primed trials. The information of all subject were combined in a single prediction to obtain 72, 60, and 56 percent of decoding accuracy for *Fea*, *FS*, and *Sgram* respectively. Again, the significance of this results emerge from a Jack-knife analysis as driven by the whole distribution, as opposed by single outliers (one-way repeated measures ANOVA, $p < 0.0005$).

3.3 Temporal Response Functions

The quantitative measure provided by the EEG prediction correlations are indeed a main result provided by this experimental framework, as it allows to study and disentangle contributions to the models from different stages of the speech processing hierarchy. Furthermore, the TRF approach offers the opportunity to study the model weights returned by the linear regression, which can be interpreted as the activation of macroscopic areas of the cerebral cortex in response to continuous speech at specific delays (time-lags). In the case of the current dataset, the three conditions (**a**, **b**, **c**) can be compared in terms of response to *Env* (not shown) and to *Ph* and *Fea* (**Fig. 4A** and **4B**). The y-axis labels of both **Figure 4A** and **4B** are grouped in macroscopic categories which exhibit similar patterns of activations (Di Liberto et al., submitted).

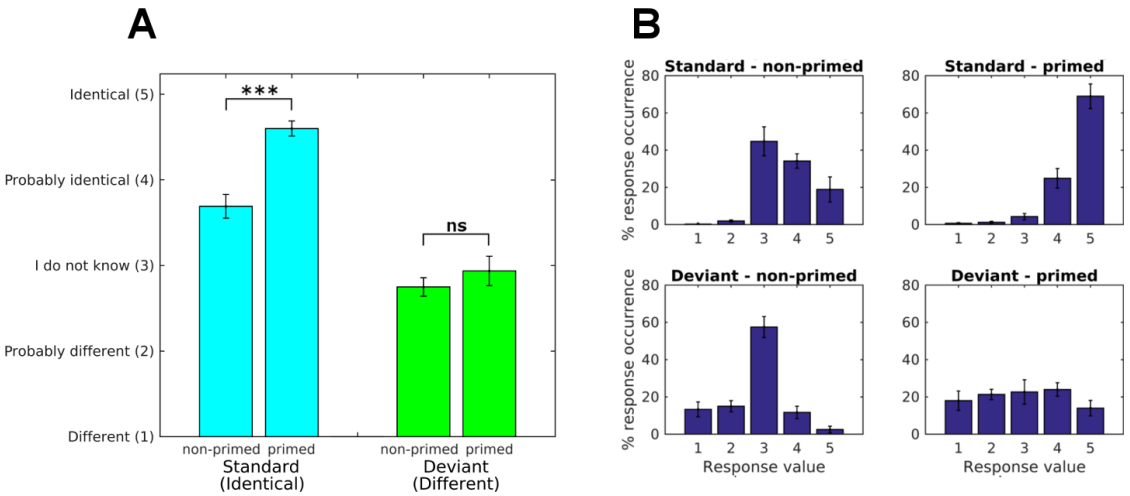


Figure 2. Subjects were presented with sequences of vocoded-original-vocoded speech snippets and were asked to identify the two noise-vocoded streams (non-primed and primed conditions) as standard or deviant presentations by comparing them with the original speech snippet. Responses consisted of a level of confidence from 1 ('different') to 5 ('identical'). **(A)**, standard/deviant identification has an interaction with the priming factor (ANOVA, $p < 0.0005$). A post-hoc *t*-test shows that the confidence in the identification of the standard trials was higher in the primed case for standard trials ($p < 0.0005$), while no significant difference emerged for the deviants ($p > 0.05$). **(B),** compares the distribution of response values across subjects in the primed and non-primed conditions, for standard and deviant trials.

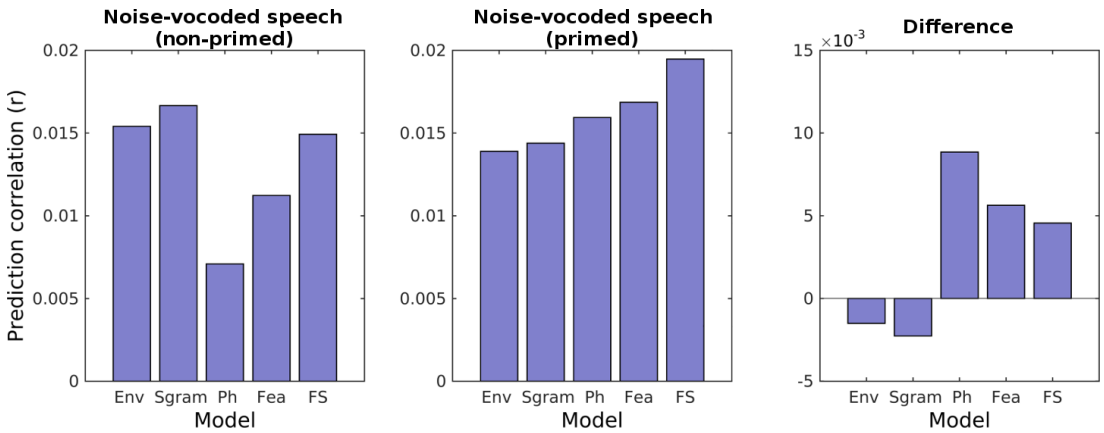


Figure 3. The models *Env*, *Sgram*, *Ph*, *Fea*, and *FS* were compared using the Pearson correlation index between the EEG signal and its prediction as a quality indicator. The results for the non-primed and primed presentations of the vocoded speech are reported in **(A)** and **(B)** respectively, while **(C)** shows their difference (**B** minus **A**), which is shown to be driven by the whole distribution, and not by single outliers ($p < 0.0005$).

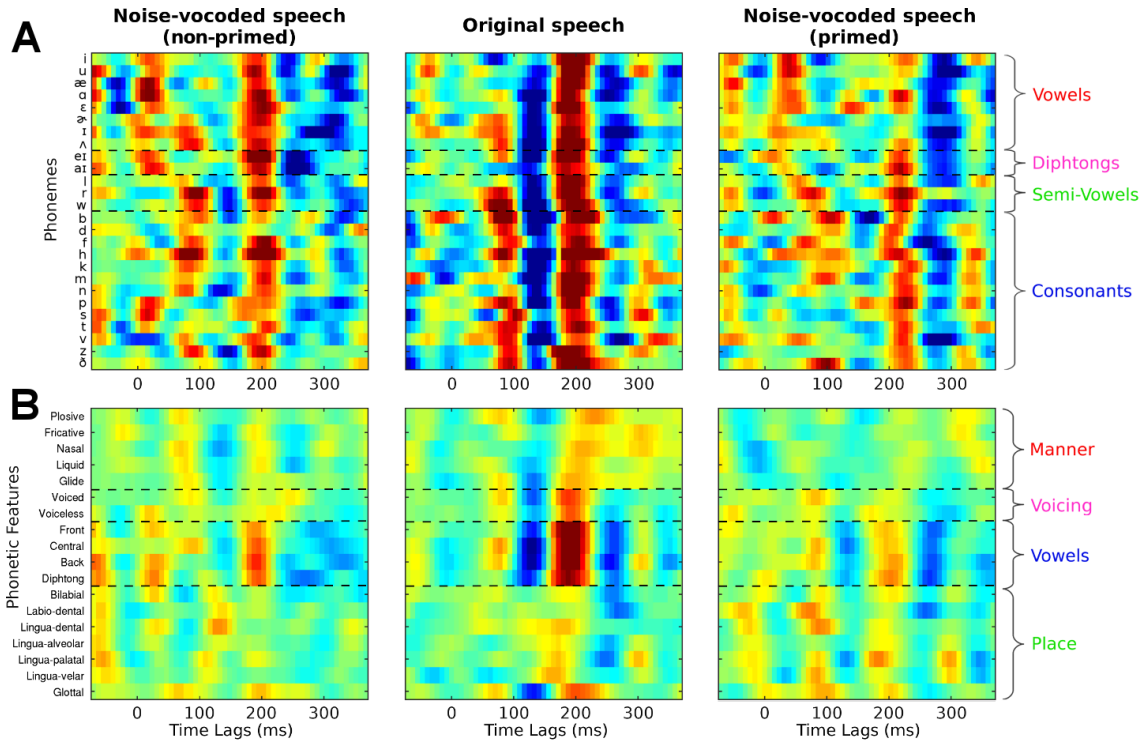


Figure 4. Temporal response functions (TRFs) in response to the phonemic model *Ph* **(A)** and to the phonetic-features model *Fea* **(B)** are compared across the clear and noise-vocoded (before and after priming) conditions. The colour-scale relates to the neural activation in response to continuous speech for a specific phoneme of phonetic-feature. The y-axis labels of both **A** and **B** are grouped in macroscopic categories which exhibit similar patterns of activations.

4 Discussion

The ability of processing natural speech correctly is of extreme importance for humans as it represents the main method of direct communication between individuals. Although social interaction can be performed in other ways, deficits in any of its aspects can be the cause of discomfort, isolation, and depression. The awareness of this issue is essential for avoiding such consequences, however the diagnosis can be problematic, especially in the less severe cases. Certainly, methodologies for studying this articulated hierarchical system have the potential to drive the understanding of new insights into its mechanisms and to serve as an instrument of diagnosis for neural disorders related to speech. In particular, we are seeking the consolidation of a dependent measure capable of quantifying the higher-level processing of speech contribution to the scalp recorded neural activity.

EEG has been shown to be sensitive to phonetic-features of speech (Di Liberto et al., submitted). These categorical features have been shown to capture information to which the envelope of speech is not sensitive, however the actual contribution of higher-level active processes related to intelligibility remains unclear. The framework we have defined here has the potential to disentangle these contributions from low-level EEG responses to spectrogram fluctuations. The results presented here support this hypothesis and motivate further studies.

The definition of a neural dependent measure of speech processing related to intelligibility could find application in clinical areas. For example, previous research (Pisoni, 2000) suggests that the inability of deaf children with cochlear implants to discriminate fine phonetic differences in place and voicing is reflected in their perception of spoken words in terms of broad phonetic categories or functional equivalence classes. Indeed, such models are potentially useful in

monitoring the development of patients in those situations and may provide new insights on the mechanisms underlying speech processing.

Acknowledgments

Funding sources: Science Foundation Ireland; Irish Research Council.

References

Aiken, S. J., & Picton, T. W. (2008). Envelope and spectral frequency-following responses to vowel sounds. *Hearing Research, 245*(1), 35-47.

Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex, 16*(1), 115-123.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. [10.1038/nn.2641]. *Nat Neurosci, 13*(11), 1428-1432.

Chomsky, N., & Halle, M. (1968). The sound pattern of English.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proc Natl Acad Sci U S A, 109*(8), E505-514. doi: 10.1073/pnas.1113427109

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (submitted). Hierarchical encoding of speech in human auditory cortex.

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America, 109*(29), 11854-11859. doi: 10.1073/pnas.1205381109

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. *The Journal of Neuroscience, 33*(4), 1417-1426.

Greenwood, D. D. (1961). Auditory Masking and the Critical Band. *The Journal of the Acoustical Society of America, 33*(4), 484-502. doi: 10.1121/1.1908699

Hickok, G. (2012). The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. *J Commun Disord, 45*(6), 393-402. doi: 10.1016/j.jcomdis.2012.06.004

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party". *Journal of Neuroscience, 30*(2), 620-628. doi: 10.1523/jneurosci.3631-09.2010

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience, 31*(1), 189-193. doi: 10.1111/j.1460-9568.2009.07055.x

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006-1010.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cereb Cortex, 20*(10), 2486-2495. doi: 10.1093/cercor/bhp318

Peelle, J. E., Johnsrude, I. S., & Davis, M. H. (2010). Hierarchical Processing for Speech in Human Auditory Cortex and Beyond. *Frontiers in Human Neuroscience, 4*, 51. doi: 10.3389/fnhum.2010.00051

Pisoni, D. B. (2000). Cognitive Factors and Cochlear Implants: Some Thoughts on Perception, Learning, and Memory in Speech Perception. *Ear and Hearing, 21*(1), 70-78.

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience, 35*(9), 1497-1503. doi: 10.1111/j.1460-9568.2012.08060.x

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America, 123*(5), 3878.

Chapter 36

Brainstem Coding of Pitch – the Butte Hypothesis

Philip X. Joris
KU Leuven, Leuven, Belgium

Contact author: Philip.Joris@med.kuleuven.be

Abstract

Pitch has been described as "the heart of hearing theory" (Plomp 2002), but its physiological underpinnings are unclear and have been much debated in preceding ISH meetings. I propose a new view in which a brainstem property plays a key role in the coding of periodicity. I refer to this property as "entracking", a contraction of entrained phase-locking. A robust code for pitch exists in the auditory nerve in the form of a pooled interspike interval (ISI) distribution (Cariani and Delgutte 1996), which resembles the stimulus autocorrelation function. An unsolved question is how such a representation could be grounded in physiology, as its computation seems to require a combination of coincidence detectors and a range of delays extending over many milliseconds. I propose a scalar rather than vector code of periodicity by virtue of coincidence detectors that code the dominant ISI directly into spike rate through entracking. Perfect entracking means that a neuron fires one spike per repetition period, so that the firing rate equals the repetition frequency. Key properties of entracking are that it improves with increasing SPL and that the number of entracking neurons increases with SPL. The main limitation in this code is the upper limit of firing (usually near 500 Hz). It is proposed that entracking provides a periodicity tag which is superimposed on a tonotopic analysis. At low SPLs and fundamental frequencies > 500 Hz, a spectral or place mechanism codes for pitch. With increasing SPL the place code degrades but entracking improves and first occurs in the neurons with the lowest thresholds for the spectral components present. The prediction is that populations of neurons extended across CF form plateaus ("buttes") of firing equaling the fundamental frequency. The core idea is that entracking brings a stratification to firing rate tied to periodicity.

Manuscript not available at the time of preprint.

Chapter 37

Can Temporal Fine Structure and Temporal Envelope be Considered Independently for Pitch Perception?

Nicolas Grimault

Centre de Recherche en Neurosciences de Lyon, CNRS UMR 5292, Université Lyon 1, Lyon, France

Contact: NG, nicolas.grimault@cnrs.fr

Abstract

In psychoacoustics, works on pitch perception attempt to distinguish between envelope and fine structure cues that are generally viewed as independent and separated using a Hilbert transform. To empirically distinguish between envelope and fine structure cues in pitch perception experiments, a dedicated signal has been proposed. This signal is an unresolved harmonic complex tones with all harmonics shifted by the same amount of Hz. As the frequency distance between adjacent components is regular and identical than in the original harmonic complex tone, such a signal has the same envelope but a different fine structure. So, any perceptual difference between these signals is interpreted as a fine structure based percept. Here, as illustrated by very basic simulations, I suggest that this orthogonal point of view that is generally admitted could be a conceptual error. In fact, neither the fine structure nor the envelope is required to be fully encoded to explain pitch perception. Sufficient information is conveyed by the peaks in the fine structure that are located nearby a maximum of the envelope. Envelope and fine structure could then be in perpetual interaction and the pitch would be conveyed by "the fine structure under envelope". Moreover, as the temporal delay between peaks of interest is rather longer than the delay between two adjacent peaks of the fine structure, such a mechanism would be much less constrained by the phase locking limitation of the auditory system. Several data from the literature are discussed from this new conceptual point of view.

Keywords: Pitch, envelope, fine structure.

1 Introduction

Numerous works about pitch perception attempt to identify and to distinguish between relevant temporal cues as envelope and fine structure. These cues are generally viewed as independent. In fact, mathematically, these cues are orthogonal and can be extracted and separated using a Hilbert transform. Using this mathematical decomposition, some works suggest that musical pitch would rely mostly on the fine structure and that speech perception would rely mostly on envelope (Smith et al., 2002). To empirically distinguish between envelope and fine structure cues in pitch perception experiments, a dedicated signal has been proposed as early as 1956 (De Boer, 1956). This signal is equivalent of an unresolved harmonic complex tone with all harmonics shifted by the same amount of Hz. As the frequency distance between adjacent components of such a signal is regular and identical than in the original harmonic complex tone,

such a signal has the same envelope but a different fine structure. As a consequence, any perceptual difference between the harmonic and the shifted complex is often interpreted as a pure fine structure based percept.

Figure 1 represents the temporal information that potentially convey some pitch related temporal information. These temporal cues can be classified in three categories of periodicities. First, the periodicities of the carrier related to the delay dT1 in Figure 1. This periodicity provides some information about the frequency of the pure tone or about the frequency of the carrier. To extract dT1 periodicities, the auditory system is required to be able to phase lock on the instantaneous phase of the signal. dT1 periodicities will be called temporal fine structure periodicities (TSF_{period}) in the following. Second, the periodicities of the envelope (delay dT2 in Figure 1) provides some information about the F0 for harmonic complex tones. dT2 periodicities will be called temporal envelope periodicities (TE_{period}) in the following. Third, the periodicities of the fine structure modulated by the envelope (delay dT3 in Figure 1). This type of periodicities corresponds to the time delays between two energetic phases of the signal located nearby two successive maxima in the envelope. For harmonic complex tones, dT2 is equal to dT3 but for shifted complex tones, dT2 and dT3 are different. As dT3 is necessary a multiple of dT1, for a shifted complex, dT3 will be either equal to $n\times dT1$ or $(n+1)\times dT1$ choosing the integer n to verify the chain of inequations: $n\times dT1 \leq dT2 < (n+1)\times dT1$. As such, dT3 depends on both dT1 (fine structure information) and dT2 (envelope information) and can then be described as a periodicity related to an interaction between fine structure and envelope. dT3 periodicities will be called interaction periodicities ($TE\times TSF_{period}$) in the following.

However, most of the time, when a harmonic complex tone and a shifted complex tone with the same envelope frequency conduct to a different pitch percept, the pitch percept is supposed to be elicited by fine structure only and to be independent of envelope. This can appear as trivial but the aim of this study is simply to convince the reader that this assumption is not true.

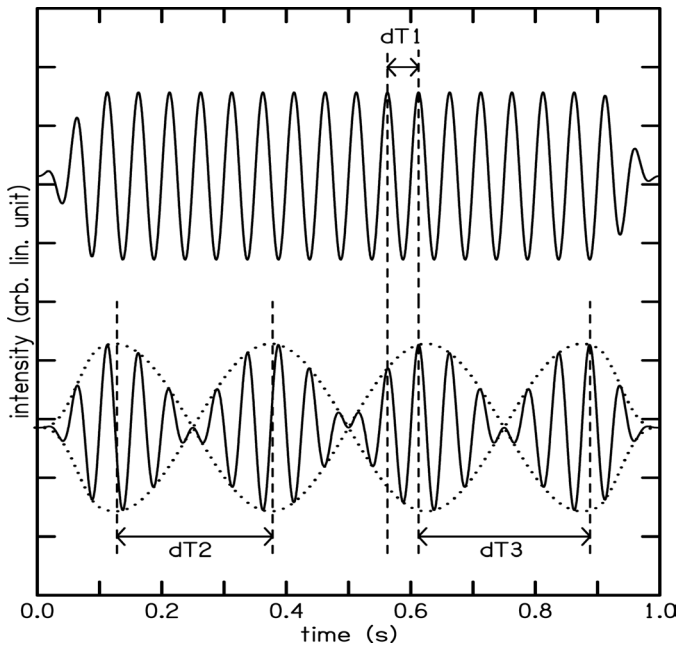


Figure 1. Potential temporal periodicity cues usable for pith perception of a pure tone (upper waveform) or a shifted complex tone (lower waveform). dT1 is the time delay between two adjacent peaks in the temporal fine structure, dT2 is the time delay between two adjacent peaks in the temporal envelope and dT3 is the time delay between two maximums of the temporal fine structure located nearby two successive maximums of the temporal envelope. dT2 and dT3 are confounded for harmonic complex tones but different for shifted complex tones.

2 Methods

2.1 Simulation

A very basic simulation has been performed to check for the potential use of various types of periodicity cues (i.e. TSF_{period}, TE_{period} or $TE \times TSF_{period}$) with various signals used in the literature. The basic idea was to pass the stimulus through a dynamic compressive gammachirp (the default auditory filter-bank of the Auditory Image Model) (Irino & Patterson, 2006) and to transform the output waveform into a spiketrain by using a threshold dependent model. This model has a higher probability to spike each time the output is over a threshold value (see equation 1). Moreover, a reasonable refractory period of 2500 Hz is added to the simulation. The periodicities of the spiketrain are finally extracted with an autocorrelation. So, the spiketrain generation is based on the following formula which depends on both envelope and fine structure temporal information:

$$\text{spiketrain}(t) = \begin{cases} 1 \text{ if } U(t) \times TFS_{signal>0}(t) \times E(t) \times P_{refract}(t) > Thres \\ 0 \text{ if } U(t) \times TFS_{signal>0}(t) \times E(t) \times P_{refract}(t) < Thres \end{cases} \quad (1)$$

where:

$U(t)$ is an uniform intern noise between 0 and 1,

$TSF_{signal>0}(t)$ is the positive part of the fine structure at the output of an auditory filter,

$E(t)$ is the envelop at the output of an auditory filter,

$P_{refract}(t)$ is the refractory period. This function is either equal to 0 or 1 related to a refractory period equal to $1/2500$. So, i*f* $spiketrain(t_0)=1$, $P_{refract}(t)=0$ *for* $t_0<t<t_0+1/2500$ *and* $P_{refract}(t_0+1/2500)=1$,

Thres is the discharge threshold sets here to 0.5,

Which is exactly equivalent to:

$$\text{spiketrain}(t) = \begin{cases} 1 \text{ if } U(t) \times Signal_{signal>0}(t) \times P_{refract}(t) > Thres \\ 0 \text{ if } U(t) \times Signal_{signal>0}(t) \times P_{refract}(t) < Thres \end{cases} \quad (2)$$

where

$Signal_{signal>0}(t)$ is the positive part of the input signal at the output of an auditory filter.

As an intern noise *U* has been added, each signal is passed 300 times in the simulation model to estimate the distribution of the periodicities of the spiketrain. Moreover, for each signal, a single auditory filter output, located in the passband of the input signal and centred on the carrier frequency (f_c), has been used.

2.2 Stimuli

The stimuli used by Santurette & Dau (2011) and by Oxenham et al. (2011) have been generated and processed through the simulation. In Santurette & Dau (2011), the signals were generated by multiplying a pure tone carrier with frequency f_c with a half-wave rectified sine wave modulator with modulation frequency f_{env} and low-pass filtered by a 4th order Butterworth filter with cut-off frequency of $0.2 \times f_c$. All signals were generated at 50 dB SPL and mixed with a TEN noise at 34 dB SPL per ERB. All f_c and f_{env} values are indicated in Figure 2. When f_c and f_{env} are not multiple from each other, this manipulation produce a shifted complex. In Oxenham et al. (2011), harmonic complex tones at various F0 values (indicated in Figure 3) were generated by adding in random phase up to 12 consecutive harmonics, beginning on the sixth. Harmonics above 20kHz were not generated. All harmonics were generated at 55 dB SPL per component and all signals were embedded in a broadband TEN noise at 45 dB per ERB. A shifted version of

each harmonic complex tone was also generated by shifting all components of the complex tone by an amount of 0.5×F0.

3 Results

The outputs of the simulation provide the distributions of the temporal periodicities of the spiketrain. This is supposed to predict the perceived pitch evoked by the signal.

The results of the simulation plotted in Figure 2 evidence that, as in Santurette & Dau (2011), the predicted pitch values are always related to $TE \times TSF_{period}$ and never related to TE_{period}. This is strongly consistent with the results reported by the authors.

The results of the simulation plotted in Figure 3 are strongly consistent with Oxenham et al. (2011). Using a plausible refractory period of 2500 Hz, the simulation is able to extract the $TE \times TSF_{period}$ even if the TSF_{period} are too fast to be correctly encoded. The decrease in performances reported in Oxenham et al. (2011) when increasing the F0 from 1200 Hz to up to 2000 Hz is also simulated. This decrease is probably explained by a decrease in resolvability (less and less interactions between harmonic components) when increasing the F0.

Finally, the results of the simulation plotted on the right column of Figure 4 are strongly consistent with the results found with the signals used by Santurette & Dau (2011), and also evidence that the predicted pitch values are always related to $TE \times TSF_{period}$ and never related to TE_{period}.

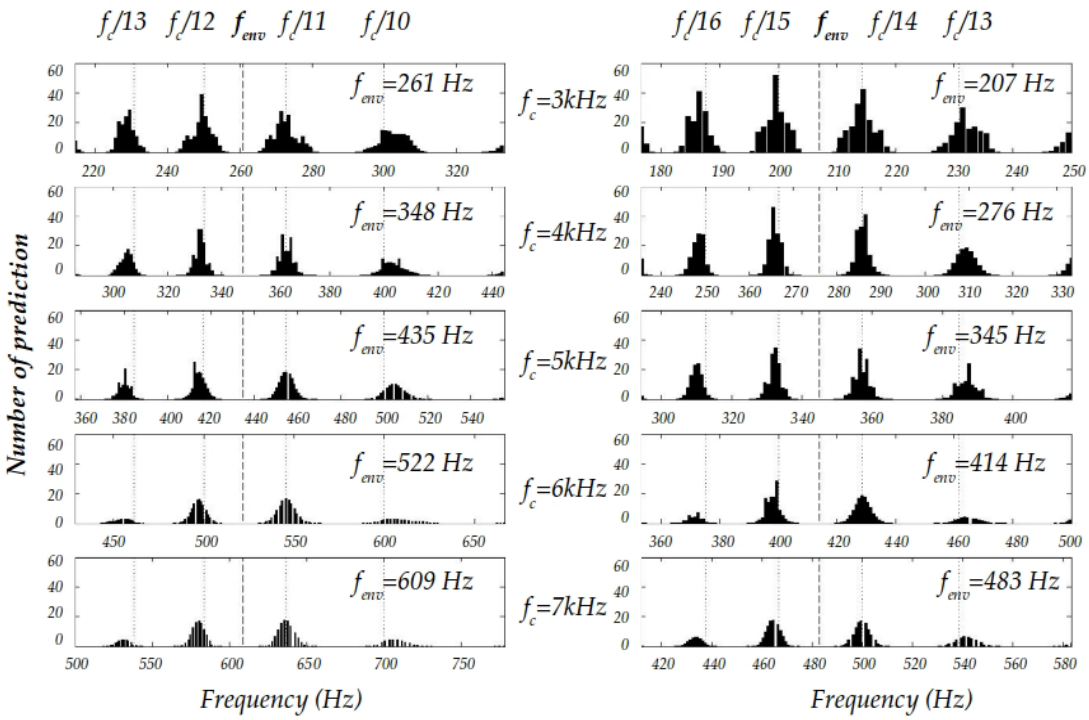


Figure 2. Outputs of the simulation fed with the signals used in Santurette & Dau (2011). The distributions of pitch estimation are always related to $TE \times TSF_{period}$ and never on TE_{period}. This is closely consistent with the data reported in Figures 4 and 6 in Santurette & Dau (2011).

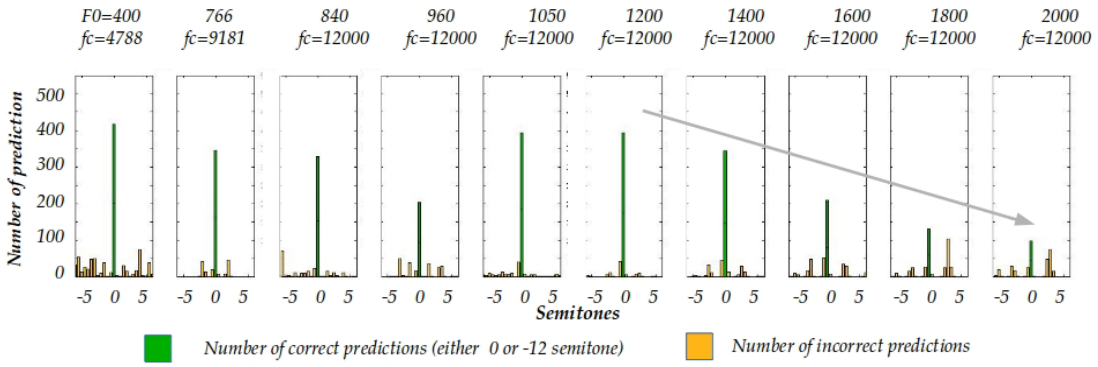


Figure 3. Outputs of the simulation at the auditory filters centred on fc, fed with the signals used in Oxenham et al. (2011). The distributions of pitch estimation are closely consistent with the data reported in Figure 2B in Oxenham et al. (2011). The gray arrow indicates the decrease in the number of predictions when increasing the resolvability of the complex by increasing the F0.

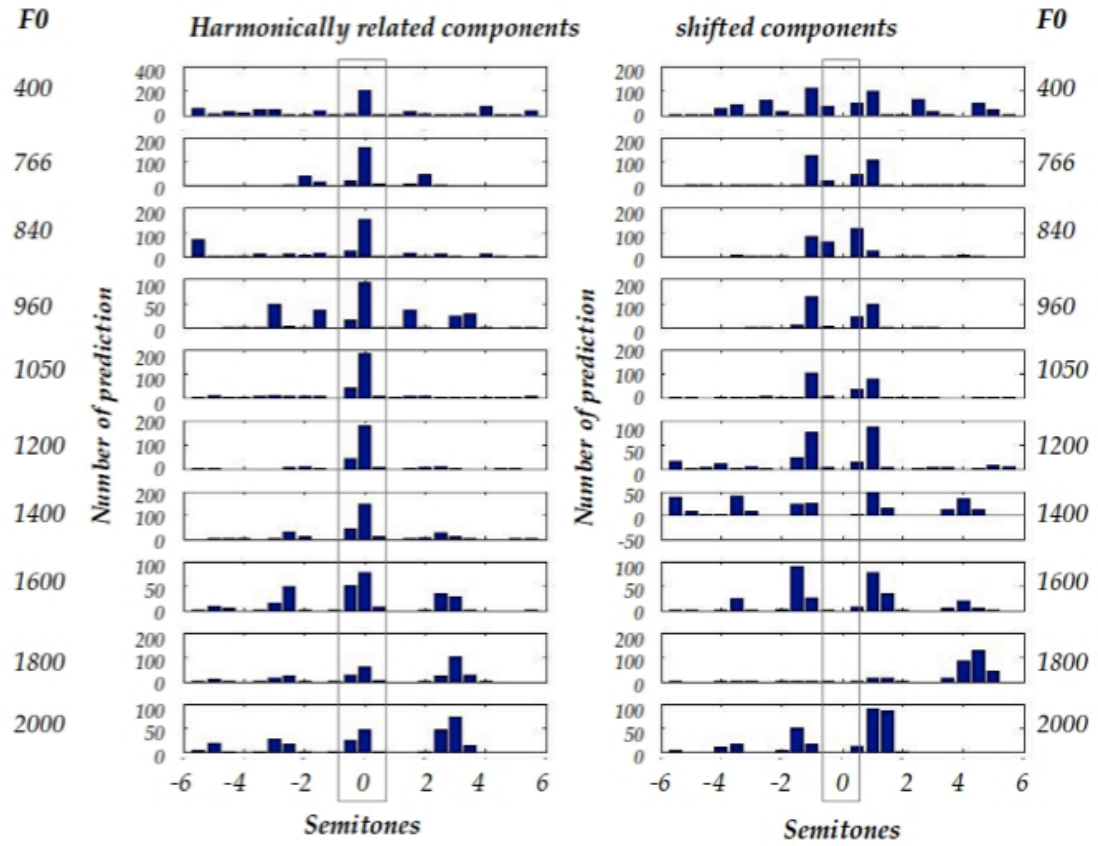


Figure 4. Output of the simulation fed with the signals used in Oxenham et al. (2011) with harmonic complex tones (left column) and shifted complex tones (right column). The distributions of pitch estimation are always related to $TE \times TSF_{period}$ which predict a peak centred in the gray rectangle on the left and peaks on either side of the gray rectangle on the right. This is closely consistent with the data reported in experiment 1 and 2 in Oxenham et al. (2011).

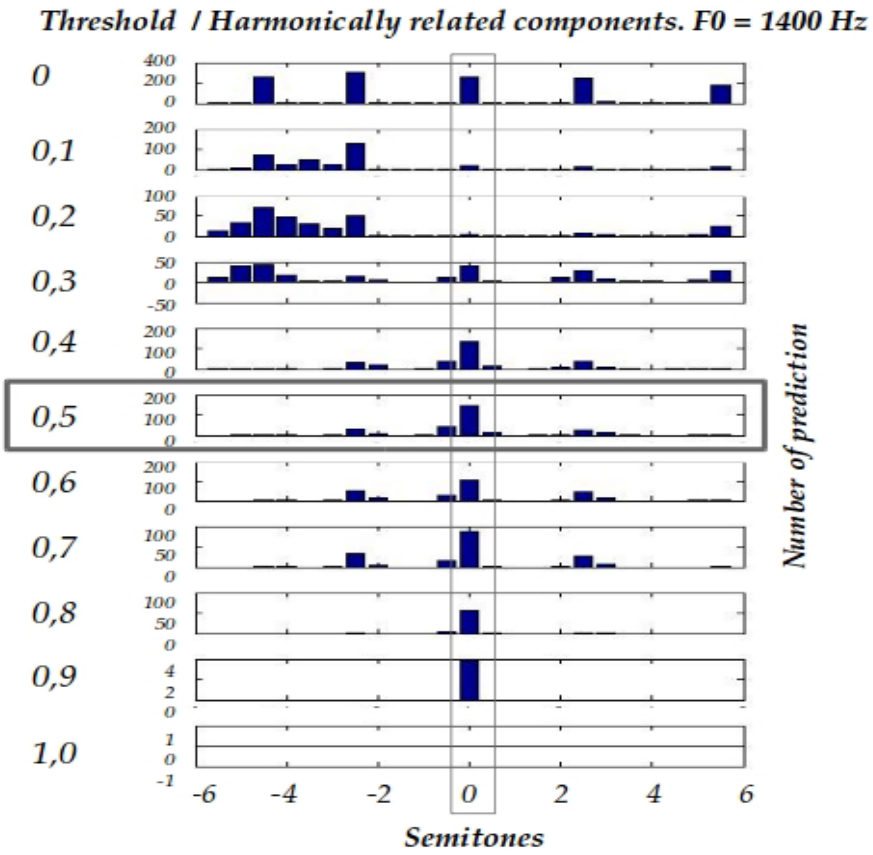


Figure 5. Output of the simulation fed with a single harmonic complex tone (F0=1400 Hz) used in Oxenham et al. (2011). Effect of threshold value from 0 to 1 on periodicity estimations.

As a control, the effect of threshold value used in the simulation has been tested with one complex tone having a F0 equal to 1400Hz (Figure 5). Varying the threshold value have some important incidence on the predictions. Using a threshold below 0.3 does not provide reliable periodicities estimations. Using a threshold from 0.4 to 0.8 provides reliable and consistent estimations. Using a high threshold value (over 0.9) prevents to report any periodicities. Using 0.5 as in the previous simulations appears then to be a good compromise. It is worth noting that such a threshold model is physiologically plausible and could be related to the thresholds of the auditory nerve fibres previously described in the literature (Sachs & Abbas, 1974).

4 Discussion

4.1 Conclusions

These stimulations have a double interest.

First, this evidence that any perceptual effect that is empirically evidenced between harmonic complex tones and shifted complex tones should not been interpreted as a pure effect of fine structure. In fact, the pitch evoked by a shifted complex is based on interaction cues between envelope and fine structure ($TE \times TSF_{period}$). Using these signals to tease apart temporal envelope cues from temporal fine structure cues is then a conceptual error. This impaired the conclusion that the pitch of unresolved complex tones is based *only* on fine structure information.

Second, when thinking about pitch perception of unresolved complex tones in terms of interaction between envelope and fine structure, it appears that the limitation of phase locking is probably much less critical than when thinking in terms of fine structure only. In fact, it seems clear that there is no need to encode every phases of the signal to encode the most intense phases located nearby an envelope maximum. This explains that the simulation can extract a periodicity related to pitch when the carrier frequency is over 10kHz (Figure 3).

4.2 Limitations

First, the current simulations are not a physiologically-based model of pitch perception and the refractory period which is used here does not accurately describe the physiological constraints for the phase locking. Some further works that would use realistic models of the auditory periphery should be used for further explorations.

Second, this simulation does not explain all the data reported in the literature about pitch perception. For example, experiment 2c in Oxenham et al. (2011) reports some pitch perception using dichotic stimulations with even-numbered harmonics presented on the right ear and odd-numbered harmonics presented to the left ear. This experimental manipulation increases the resolvability of the signals and prevents our simulation to extract any temporal periodicities and then to predict some pitch perception.

Acknowledgments

This work was supported by institutional grants ("Centre Lyonnais d'Acoustique," ANR-10-LABX-60) and (Contint - Aïda, ANR-13-CORD-0001). I thank Andy Oxenham and Etienne Gaudrain for constructive and interesting comments about this work.

References

De Boer, E. (1956). Pitch of inharmonic signals. *Nature*, *178*(4532), 535-536.

Oxenham, A. J., Micheyl, C., Keebler, M. V., Loper, A., & Santurette, S. (2011). Pitch perception beyond the traditional existence region of pitch. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(18), 7629-7634. http://doi.org/10.1073/pnas.1015291108

Santurette, S., & Dau, T. (2011). The role of temporal fine structure information for the low pitch of high-frequency complex tones. *The Journal of the Acoustical Society of America*, *129*(1), 282-292. http://doi.org/10.1121/1.3518718

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87-90. http://doi.org/10.1038/416087a

Chapter 38

Locating Melody Processing Activity in Auditory Cortex with MEG

Roy D. Patterson
Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK

Martin Andermann
Section of Biomagnetism, Department of Neurology, University of Heidelberg, Heidelberg, Germany

Stefan Uppenkamp
Medical Physics Section, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Andre Rupp
Section of Biomagnetism, Department of Neurology, University of Heidelberg, Heidelberg, Germany

Contact author: RDP, rdp1@cam.ac.uk

Abstract

This paper describes a technique for isolating the brain activity associated with melodic pitch processing. The magnetoencephalograhic (MEG) response to a four note, diatonic melody built of French horn notes, is contrasted with the response to a control sequence containing four identical, "tonic" notes. The transient response (TR) to the first note of each bar is dominated by energy-onset activity; the melody processing is observed by contrasting the TRs to the remaining melodic and tonic notes of the bar (2-4). They have uniform shape within a tonic or melodic sequence which makes it possible to fit a 4-dipole model and show that there are two sources in each hemisphere – a melody source in the anterior part of Heschl's gyrus (HG) and an onset source about 10 mm posterior to it, in anterior planum temporale (PT). The N1m to the initial note has a short latency and the same magnitude for the tonic and the melodic sequences. The melody activity is distinguished by the relative sizes of the N1m and P2m components of the TRs to notes 2-4. In the anterior source a given note elicits a much larger N1m-P2m complex with a shorter latency when it is part of a melodic sequence. The TRs are accompanied by a sustained field (SF) originating from a source in HG near the melody source. However, the SF magnitude is of comparable size for the tonic and melodic bars. This study shows that melodies evoke activation in a separate, anterior part of auditory cortex.

Keywords: auditory cortex, melodic pitch, magnetoencephalography

1 Introduction

Recently, Patterson et al. (2014) extended earlier imaging research on melody processing (Patterson et al., 2002; Rupp and Uppenkamp, 2005; Uppenkamp and Rupp, 2005) by presenting listeners with four-bar musical phrases played by a French horn. There were four notes per bar; they had the same amplitude and the inter-note gaps where minimized to minimize the energy onset component of the responses to the second and succeeding notes 2-4 (Krumbholz et al.,

2003). In the first and third bars, the pitch was restricted to the tonic of the four-bar phrase; the second bar contained a novel, random tonal melody and the fourth bar contained a variation of that melody. Magnetoencephalographic recordings revealed that each note produced a prominent transient response (TR) in auditory cortex with a P1m, an N1m and a P2m. The structure of the four-bar phrase was reflected in the magnitude of the TR. In all bars, the TR to the first note was somewhat larger than that to the remaining three notes of that bar. In the melody bars, the P2m component of the TR was larger than in bars where the pitch was fixed. The stability of the TR magnitude to notes 2-4 of each bar made it possible to do a four-dipole analysis and show that the activity is distributed between two sources in each hemisphere, one in the medial part of Heschl's gyrus (HG) and the other slightly more lateral and posterior on the front edge of planum temporale (PT). The source waves showed that melody processing was largely confined to the P2m component of the TR in the medial part of HG.

This paper describes a shorter version of the melody experiment with alternating tonic and melodic bars and an inter-sequence interval of about 1325 ms – modifications that make it possible to measure the sustained fields (SFs) generated during the tonic and melodic bars. As in the previous experiment, the magnitude of the TR to the first note of the first bar is enhanced by what appears to be an energy-onset response. The purpose of this second experiment was to determine 1) whether the TR to the first note of each sequence emanates from a different source than the TRs from the remaining notes of that sequence, and 2) whether the source of the SF is in auditory cortex or beyond.

2 Methods

A natural French horn tone (Goto et al. 2013) served as the basic stimulus for the experiment. Four-note sequences were constructed by varying the f_0 of the original note within the original envelope using the vocoder STRAIGHT (Kawahara et al. 1999; Kawahara & Irino 2004). There were two different types of four-note sequences, "tonic" and "melodic": The f_0 was fixed in the tonic sequences at either 100 Hz or 141.5 Hz. In the melodic sequences, the f_0 value of each note was randomly drawn, without replacement, to be -1, 0, 2, 4, 5 or 7 semitones relative to the respective tonic value. During the MEG recordings, tonic and melodic sequences were presented in alternating order with a total of 400 sweeps each (200 sweeps per tonic). Within a given sequence, each note had a length of 650 ms and was ramped with 5-ms and 10-ms Hanning windows at the onset and offset, respectively. There were no pauses between adjacent notes within a sequence; the inter-sequence interval varied randomly between 1300-1350 ms. The total duration of the experiment was 55 minutes and the overall stimulation level was 70 dB SPL.

The neuromagnetic activity in response to the four note sequences was measured with a Neuromag-122 (Elekta Neuromag Oy) whole-head MEG system, at a 1000 Hz sampling rate, having been lowpass filtered between DC and 330 Hz. Ten normal-hearing listeners (mean age: 33.9 ± 12.01 years) participated in the study. Subjects watched a silent movie, while their AEFs were recorded; they were instructed to direct their attention to the movie. Spatio-temporal source analyses (Scherg 1990) were performed offline using the BESA 5.2 software package (BESA Software GmbH). Separate source models with one equivalent dipole in each hemisphere were created to analyze the P50m (using unfiltered data) and the sustained field (SF) (using data lowpass-filter filtered at 1 Hz). Then, a model with two dipoles per hemisphere was built to localize the N1m/P2m complex associated with energy onset at the start of each sequence, and to localize changes between adjacent notes within a given sequence (pooled notes 2-4). A 1-30 Hz bandpass filter was applied to the data for these four-dipole fits. Afterwards, unfiltered source waveforms were derived for each model for each participant to compute grand-average source waveforms for both the tonic and melodic sequences. A principal component analysis was performed on the last few milliseconds of the epoch to compensate for drift. Statistical evaluation of specific component differences was accomplished in MATLAB using the

multivariate statistic t_{sum} from the permutation test of Blair & Karniski (1993). Latencies and amplitudes of the additional data set described in the discussion section were evaluated statistically using a bootstrap technique to derive critical *t*-values (Efron and Tibshirani, 1993).

3 Results

The data were fitted using a 4-dipole model with 2 dipoles in each hemisphere. One pair focused on the N1m temporal region of the first note in each bar; the other focused on the N1m temporal regions of the three adapted melodic notes of each trial (notes 2-4) – a technique developed for analysing melody processing in Patterson et al. (2014). These melodic-sequence sources were found to be located in the medial part of Heschl's gyrus (HG); the onset-sequence sources were located about 10 mm more posterior in the anterior part of planum temporale (PT), which is a common location for the onset N1m (anterior-posterior difference, left hemisphere: $t(9)=4.21$, $p=0.0023$; right: $t(9)=4.02$, $p=0.003$).

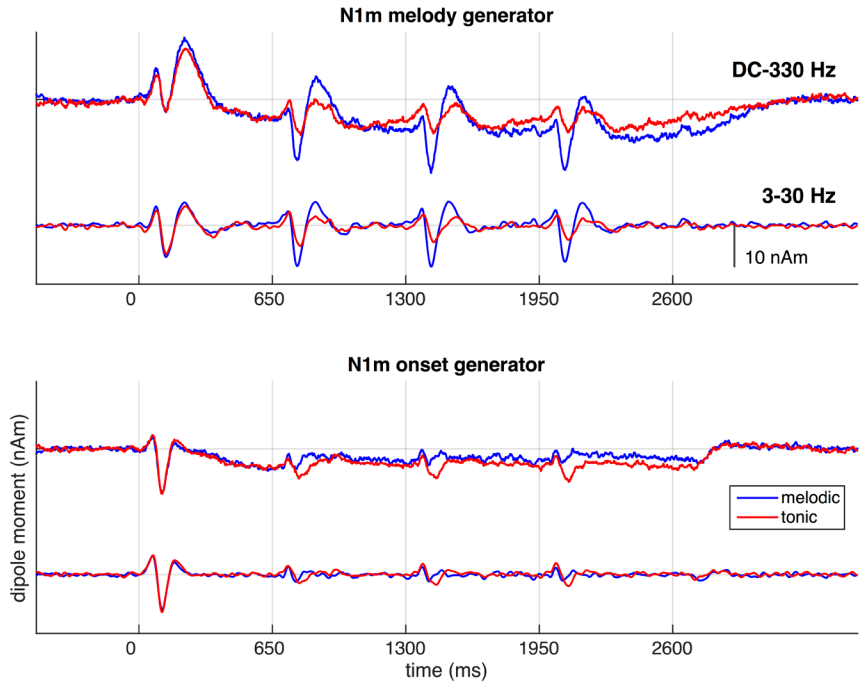


Figure 1. Average source waves extracted from the data of the tonic (red) and melodic (blue) sequences using (a) the spectro-temporal filter derived with the data from the N1m temporal region of notes 2-4 of the melodic sequences, and (b) the spectro-temporal filter derived with the data from the N1m temporal region of the initial notes of both types of sequence. The upper trace in each panel shows the unfiltered, grand average waveforms. The sustained field is removed from the data in the lower trace of each panel by band pass filtering (3-30 Hz).

The melodic-sequence sources were, then, used as spectro-temporal filters to extract source waves for the adapted notes of both the tonic sequences and the melodic sequences of each trial. The onset-note sources were similarly used to extract source waves for both the tonic and melodic sequences of each trial. The source waves for the left and right hemispheres were very similar, for both the adapted-note and the onset-note analyses, so the waves from the two hemispheres were averaged, in each case. These four average source waves are presented in Figure 1.

The top panel shows the source waves for the adapted notes in the tonic (red) and melodic (blue) sequences, superimposed for comparison. In this anterior source, the response to the first note of the tonic sequence is very similar to the first note of the melodic sequence (N100m: t_{max}=-65.13, p=0.151; P200m t_{max}=312.95, p=0.0001). Thereafter, however, the adapted responses to notes 2-4 in the melodic sequence (blue) are substantially larger than their counterparts in the tonic sequence (red), both with regard to the N1m component and the P2m component (N100m: t_{max_note2}=-204, p=0.002; t_{max_note3}=-229.35, p=0.002; t_{max_note4}=-239.27, p=0.002; P200m: t_{max_note2}=416.04, p=0.0001; t_{max_note3}=427.91, p=0.002; t_{max_note4}=443.30, p=0.002). The level, duration and timbre of the notes in the tonic and melodic sequences are the same, so the differences in the source waves reflect the processing of the sequence of pitch changes that together constitute the melody.

The data were not filtered in these analyses and the time interval between trials varied randomly between 1300 and 1350 ms in the experiment. This made it possible to extract and compare the sustained fields (SFs) produced in the tonic and melodic sequences. In this melodic source, the SF builds up (negatively) over about 500 ms following the P2m to the first note, from about 250 to 750 ms after the onset of the sequence. The magnitude of the SF decreases down to about -20 nAm below baseline for both sequences; it appears to get a little more negative towards the end of the melodic sequence, but not a great deal. Then, following the P2m of the last note, the SF decays back up to the baseline level over about 500 ms for both sequences. Statistically, over the course of the entire sequence, the SFs for the tonic and melodic conditions are not different, so it appears that the process that produces the SF does not distinguish the condition in which the notes produce a melody (SF differences based on melodic sources: -1973.99, p=0.29; SF based on the 2-dipole SF-fit (0-1 Hz): t_{max}=1236.12, p=0.545).

The bottom panel shows the source waves derived for the tonic (red) and melodic (blue) sequences with the spectro-temporal filter from the onset notes. In this posterior source, the responses to these sequence-initial notes are virtually identical (t_{max}=21.81, p=0.688). The differences between these source waves appear in the transient responses to the second and succeeding notes. The transient responses to the notes in the tonic sequence exhibit a clear, rather broad, N1m component; the transient responses to the notes of the melodic sequence are so small as to be barely detectable.

There is a SF in the posterior source waves that builds up to about -10 nAm over about 500 ms following the P2m to the first note, as in the anterior source. Over the course of the remaining notes of the sequence, the SF varies little, although the level appears to become slightly more negative towards the end of the *tonic* sequence (which is the opposite of the case in the anterior source). Then, following the P2m of the last note of both sequences, the SF decays back up to the baseline level, as with the anterior source. However, it does so much more quickly in this posterior (onset) source than in the anterior (melodic) source; it decays over about 100 ms rather than 500 ms. That said, the SFs for the tonic and melodic sequences are not significantly different, so the process that produces the SF in the posterior source does not distinguish the condition in which the notes produce a melody, just as with the anterior source (t_{max}=4249.26, p=0.058).

4 Discussion

The transients produced by the adapted notes (2-4) have consistently different shapes in the melodic and onset sources; they are also consistently larger in the melodic source. This suggests that separate components of the N1m are generated in the posterior and anterior sources. This prompted us to return to the melody data of Patterson et al. (2014) and do a more detailed analysis of the transient responses. In that version of the melody experiment, there are two tonic bars and two melodic bars in each trial, and in each case, the six adapted responses from the bars with the same sequences (tonic or melodic) are very similar. This made it possible to fit a

four-dipole model to the activity in HG with the focus either on the N1m temporal region or the P2m temporal region, which in turn, allowed us to compare the role of the N1m and P2m in the two experiments.

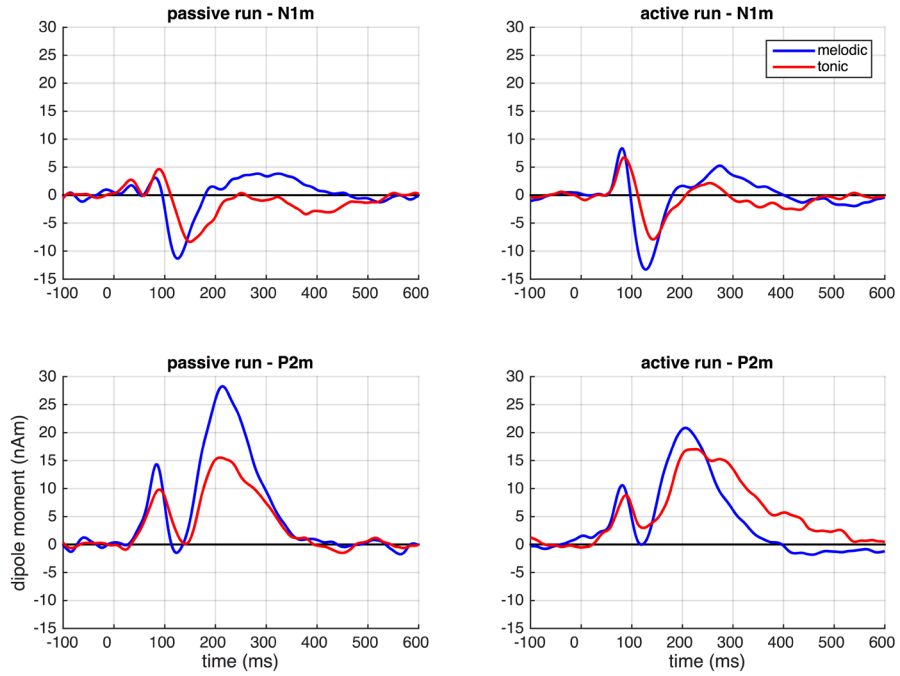


Figure 2. Average transient responses from the 4-dipole model. Top row: waves derived with the N1m temporal window. Bottom row: waves derived with the P2m temporal window. Left column: waves derived from the passive listening condition. Right column: waves derived from the active-listening condition.

Figure 2 presents a comparison of the tonic-note responses from bars 1 and 3 (red lines) with the melodic-note responses from bars 2 and 4 (blue lines). The upper row shows the TRs from the source waves obtained with the N1m temporal window; the lower row shows the TRs obtained with the P2m temporal window. In this experiment there was both a passive listening condition (left column) and an active listening condition (right column) in which the listener had to decide which note changed from the first to the second presentation of the melody. There is a P1m in both TRs of all four panels; the largest difference in magnitude is between the melody condition with the N1m temporal window (top left panel) and the melody condition with the P2m temporal window (bottom left panel). But the differences are not large leading to the common conclusion that the P1m just marks the detection of the onset of a sound.

In the upper panels, the P1m is followed by a pronounced N1m in both the tonic TR and the melodic TR, and in both the active and passive conditions. In contrast, there is essentially no P2m activity in the N1m temporal window data. The reverse is the case in the lower row: The P1m is accompanied by a pronounced P2m in both the tonic TR and the melodic TR, and in both the active and passive conditions; there is essentially no N1m activity in the P2m temporal window data.

The upper panels also show that the N1m peak occurs earlier in the *melodic* TR (blue) where one might have thought that it would take longer to produce a response (latency $\text{diff}_{\text{passive}}$ = 25.63 ms, crit-t=14.15, p<0.05; latency $\text{diff}_{\text{active}}$ = 14.22 ms, crit-t=8.21, p<0.05). Although the peak delay differences might seem small, the timing of the N1m peak is one of the most precise statistics in MEG data, and these differences are highly significant – more so than the N1m

amplitude differences in these panels (amplitude $diff_{passive}$ = 2.74 nAm, crit-*t*=2.6, n.s.; amplitude $diff_{active}$ = 5.52 nAm, crit-*t*=5.35, $p<0.05$).

The lower panels show that, in the passive listening condition, the melodic-note sequences produce a larger P2m than the tonic-note sequences (left panel; amplitude $diff_{passive}$ = 12.52 nAm, crit-*t*=4.57, $p<0.05$), but the difference is less in the active condition (right panel; amplitude $diff_{active}$ = 3.72 nAm, crit-*t*=4.70, n.s.), both because the melody sequence response decreases and the tonic sequence response increases at longer delays.

5 Conclusions

The results of the current experiment confirm that the energy onset component of the N1m arises in a generator in PT posterior to the melody processing generator which is in the anterior part of HG. There is a melody onset component in the response to the first note of melodic sequences, which is confined to the P2m. The sustained field originates from a generator in HG not far from the melody generator; however, the SF itself does not distinguish between the tonic and melodic sequences. Finally, a reanalysis of the TR data from the previous experiment suggests that the N1m and P2m components of the melodic response in HG can be separated with a 4-dipole source model, and that melody processing activity is more closely associated with the P2m component of the response.

References

Blair, R.C. & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology* 30, 518–524.

Efron, B., & Tibshirani, R.J. (1993). An Introduction into the Bootstrap. New York: Chapman & Hall.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2003). RWC music database: music genre database and musical instrument sound database. In: Hoos, H.H., Bainbridge, D. (Eds.), *Proceedings of the 4th International Conference on Music Information Retrieval*, pp. 229–230.

Kawahara, H. & Irino, T. (2004). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In:Divenyi, P.L. (Ed.), *Speech Separation by Humans and Machines*. Kluwer Academic, Dordrecht, pp. 167–180.

Kawahara, H., Masuda-Kasuse, I. & de Cheveigne, A. (1999). Restructuring speech representations using pitch-adaptive time-frequency smoothing and instantaneous-frequency based F0 extraction: possible role of repetitive structure in sounds. *Speech Commun*. 27, 187–204.

Krumbholz, K., Patterson, R. D., Seither-Preisler, A., Lammertmann, C. & Lütkenhöner, B. (2003). Neuromagnetic evidence for a pitch processing centre in Heschl's gyrus. *Cerebral Cortex* 13 765-772.

Patterson, R. D., Uppenkamp, S., Andermann, M. & Rupp, A. (2014). "Brain imaging the activity associated with pitch intervals in a melody," *ASA meeting in Providence, Rhode Island*, 5-9 May 2014.

Patterson, R.D., Uppenkamp, S., Johnsrude, I. & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron* 36 767-776.

Rupp, A. & Uppenkamp, S. (2005). Neuromagnetic representation of short melodies in the auditory cortex. *Fortschritte der Akustik-DAGA 2005*. Eds Hugo Fastl & Markus Fruhmann, DEGA e.V., Berlin, 473-474.

Scherg, M. (1990). Fundamentals of dipole source potential analysis. In: Grandori F., Hoke M., Romani G.L. (eds), Auditory evoked magnetic fields and potentials. *Advances in Audiology*, vol. 5. Karger, Basel, pp. 40-69.

Uppenkamp, S. & Rupp, A. (2005). Functional MR imaging of the processing of pitch changes in human listeners. *Fortschritte der Akustik-DAGA 2005*. Eds Hugo Fastl & Markus Fruhmann, DEGA e.V., Berlin, 471-472.

Chapter 39

Studying Effects of Transcranial Alternating Current Stimulation on Hearing and Auditory Scene Analysis

Lars Riecke

Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

Contact: L.Riecke@MaastrichtUniversity.NL

Abstract

Recent studies have shown that perceptual detection of near-threshold auditory events may depend on the timing of the event, relative to the phase of ongoing brain oscillations (i.e., periodic fluctuations in neural excitability). It has been further shown that transcranial alternating current stimulation (tACS), a non-invasive and silent brain stimulation technique, can entrain ongoing cortical alpha oscillations and thereby offer some experimental control over the phase of these oscillations. Based on these and related alpha-tACS findings, the present study investigates the potential of delta/theta-tACS to modulate hearing and auditory scene analysis in normally-hearing listeners. Detection performance is measured in silence or in a multi-tone background for near-threshold auditory stimuli that are modulated at 4Hz and presented at various moments (phase lags) during ongoing oscillatory electric stimulation (two synchronous 4-Hz alternating currents applied transcranially and simultaneously to the two cerebral hemispheres). Preliminary results indicate that performance fluctuates as a function of phase lag, and these fluctuations can be explained best by a sinusoid whose frequency matches that of the tACS. This suggests that tACS may amplify/attenuate sounds that are temporally coherent/anticoherent with it and thereby enhance/reduce their audibility.

Keywords: brain stimulation, oscillation, entrainment, delta, theta, auditory cortex

1 Introduction

1.1 What Is TACS?

TACS is a non-invasive brain stimulation technique. It involves inducing a weak alternating electric current between electrodes at the scalp (Nitsche et al., 2008; Zaghi et al., 2010; Paulus, 2011). The current partially penetrates the brain, where it propagates widely and forces neuronal excitability to oscillate along with the alternating current waveform (Frohlich & McCormick, 2010; Ozen et al., 2010). This way, alternating currents at a frequency between 0.5-20Hz may entrain ongoing cortical oscillations (Marshall et al., 2006; Pogosyan et al., 2009; Zaehle et al., 2010; Neuling et al., 2012; Ali et al., 2013; Neuling et al., 2013; Helfrich et al., 2014) with the same frequency. Depending on stimulation parameters such as electrode positions, this tACS-induced cortical entrainment can impact on different aspects of sensory-perceptual-motor processing (phosphenes, visual contrast, visual orientation, visual stability, tone audibility,

tactile sensitivity) or higher-order cognition (memory, decision-making), as summarized elsewhere (Herrmann et al., 2013). TACS can be applied safely and silently using portable and relatively inexpensive equipment. While neural entrainment is currently a popular topic in the auditory literature (see next section), auditory tACS studies are surprisingly scarce so far (Neuling et al., 2012).

1.2 Cortical Oscillatory Phase Influences Auditory Perception

Cortical oscillations may entrain to a periodic auditory stimulus so that phases of high neuronal excitability become temporally aligned with peaks in the stimulus envelope. This sound-brain entrainment is thought to facilitate cortical processing of upcoming acoustic events that coincide with the peaks in the sound envelope, as these events will arrive during phases of relatively high neuronal excitability (Schroeder & Lakatos, 2009; Arnal & Giraud, 2012). Although several correlational studies have shown that sound-brain entrainment plays an important role for the neural processing and perception of auditory stimuli (Stefanics et al., 2010; Besle et al., 2011; Gomez-Ramirez et al., 2011; Lakatos et al., 2013), only three published human studies so far have sought to directly control this entrainment with periodic external stimulation and assess the perceptual effects. Neuling et al. (2012) presented a tone pip in noise and simultaneously applied 10-Hz tACS above the auditory cortices while manipulating the timing of the target tone relative to the phase of the ongoing tACS (i.e., the phase lag). They found that detection performance varied as a function of phase lag and that this function resembled a 10-Hz sinusoid, i.e., the tACS waveform. They concluded that 10-Hz tACS can entrain cortical alpha oscillations and thereby modulate auditory perception in a phase-specific manner. Instead of tACS, Henry et al. (2012; 2014) used a periodic 3- or 5-Hz frequency modulation, applied to an ongoing tone, as entraining stimulus. They manipulated the timing of a gap in the tone relative to the phase of the ongoing frequency modulation and found that this influenced gap detection in the same phase-dependent manner as described above for the tACS study. Based on electrophysiological recordings (Lakatos et al., 2005; Frohlich & McCormick, 2010; Ozen et al., 2010; Ali et al., 2013) and findings from visual entrainment studies (Romei et al., 2012; de Graaf et al., 2013; Jaegle & Ro, 2014; Spaak et al., 2014), it appears plausible that the reported effects on auditory perception were mediated by cortical oscillations that entrained to the periodic external stimulation and thereby caused the auditory targets to arrive at phases of different neural excitabilities. Together these neuromodulation findings suggest that periodic electric or acoustic stimulation offer some control over the timing of cortical excitability in the alpha or delta/theta range and therewith over the detectability of brief acoustic events that occur during the periodic stimulation.

1.3 Can TACS Influence Auditory Perception?

Compared with periodic acoustic stimulation, tACS arguably provides the preferred entrainment stimulus for auditory studies as it can be applied both in silence and at selected scalp position, thus allowing bypassing the peripheral auditory system and reducing possible masking of the target sound. However, some authors have argued that the observed effects of tACS on auditory perception could merely reflect middle-ear muscle contractions induced by current leakage from central to peripheral auditory system (Zoefel & Heil, 2013) or more generally have questioned the capability of tACS to entrain brain oscillations (Brignani et al., 2013).

Thus, several important issues remain unresolved: How reliable are auditory effects of tACS? Do they also apply to oscillatory frequencies outside the alpha range? Do they also occur in more naturalistic situations where target sounds need to be tracked over time amidst other sounds?

The present manuscript provides preliminary data from ongoing tACS research that aims at answering some of these questions. A first experiment investigated the reliability of tACS effects phase on sound detection by aiming to replicate the results by Neuling et al. (2012) using a more rigorous methodological approach (Riecke et al., under review). A second ongoing experiment is using an informational masking paradigm to explore whether the previous results based on tone

pips can be extended to more sustained, modulated target sounds in auditory scenes. TACS is applied throughout in the delta/theta range. Amplitude fluctuations in speech signals in this range have been shown to be critical for speech comprehension (Shannon et al., 1995) and entrain oscillations in human auditory cortex (Luo & Poeppel, 2007; Kerlin et al., 2010; Ding & Simon, 2013; Peelle et al., 2013; Zion Golumbic et al., 2013). Moreover, sounds modulated at these frequencies are highly effective in exciting neurons in the human middle superior temporal gyrus and lateral Heschl's gyrus in both cerebral hemispheres (Liegeois-Chauvel et al., 2004). Thus, these bilateral auditory cortical regions were chosen as target regions in the present experiments. Specifically, a tACS frequency of 4Hz was applied because it is well separated from, and harmonically unrelated to, the previously explored 10-Hz frequency in the alpha range.

2 Methods

2.1 Stimuli

TACS

Neuling et al. (2012) placed two tACS electrodes at positions T7 and T8 (above the temporal lobes). While this configuration may produce relatively strong currents in the auditory cortices, as suggested by electric field simulations, these currents likely propagate in opposite directions (i.e., have opposite phases) within bilateral cortical structures. In the present experiments, this possible limitation was reduced by using a novel dual-channel tACS approach. Prior electric field simulations on a standard head model using the Matlab toolbox COMETS (Jung et al., 2013) suggested that placing the stimulation electrodes (size: 5×5cm) at position T7 and T8, and the return electrodes (size: 5×7cm) close to position Cz (on the vertex), would produce relatively strong currents in the target regions in both auditory cortices. To facilitate homophasic stimulation in both auditory cortices, two approximately equivalent circuits were generated using two stimulator systems (Neuroconn, Ilmenau, Germany). Specifically, the same sinusoidal current (frequency: 4Hz, starting phase: 0°) was applied to each circuit, the return electrodes were coupled, and the skin was prepared so that the impedances of the left-lateralized and right-lateralized circuit were matched, while keeping the net impedance below 10kΩ. Current strength was set individually near threshold by reducing the peak amplitude simultaneously for both circuits in 0.1-mA steps from 1mA to the point where participants reported being uncertain about receiving tACS under every electrode. Current density did not exceed 0.04 or 0.0286mA/cm^2 for the stimulation electrodes and return electrodes respectively.

Auditory stimuli

In both experiments, the target sounds were modulated at 4Hz so that they had a similar amplitude envelope as the tACS. This gave listeners the opportunity to sample the auditory target at a fixed phase angle on consecutive tACS cycles. The relative level of the target was set individually to a value near detection threshold determined before the tACS experiment. The auditory stimuli were sampled at a rate of 16kHz and presented diotically via a soundcard (Focusrite Forte) and insert earphones (EARTone 3A) in a sound-attenuating booth.

In experiment 1, the target sound was a click train. It was generated by summing all harmonics (sinusoids with fixed amplitude and starting phase) of a 4-Hz fundamental frequency within the range from 112-3976Hz and then bandpass-filtering the resulting waveform between 224 and 1988Hz (3dB cut-off frequencies, fourth-order Butterworth filter with zero phase shift). The portion between 125 and 1125ms was extracted to obtain a train of four clicks centered on a 1-s interval.

In experiment 2, the target sound was a pulsed complex tone embedded in a continuous random multi-tone masker; for similar stimuli see (Gutschalk et al., 2008; Teki et al., 2011). The complex

tone was generated by summing three harmonics (sinusoids with fixed amplitude and starting phase) of a 110-Hz fundamental that were chosen randomly from the range from 220-6270Hz with the constraint that they differed by at least one critical band. The target was generated by modulating the complex tone with a square wave function (frequency: 4Hz, duty cycle: 16.7% corresponding to 41.7ms, starting phase: 0°). The duration of the target was chosen randomly from the interval from 2-6s. The masker was a long sequence of random complex tones, each of which was generated by summing a fixed number of sinusoids (three to five, depending on the listener's target detection threshold). The sinusoids had random starting phase and the same amplitude and duration as the individual target components. Their frequency was chosen randomly from the range from 200-6297Hz with the constraint that they differed by at least one critical band from each other and the target components (when the target was present). For intervals where a target was present, the number of masker components was reduced by three (i.e., the number of target components), so that listeners could not detect the target based on amplitude cues in the overall stimulus waveform.

2.2 Design and task

The critical parameter in both experiments was the relative timing of acoustic and electric stimulation (Fig. 1A). This parameter, which will be referred to as 'phase' for simplicity, was manipulated across six conditions. This was done by varying the onset of the modulated target in six steps of 41.7ms (30°) across the 4-Hz period of the ongoing tACS (i.e., relative to the current tACS phase angle). Each experiment lasted approximately 10min and was presented four times during tACS and once without tACS (sham stimulation). The latter was identical to tACS, except that the current was ramped down/up shortly after/before the experiment began/ended (Herrmann et al., 2013). Debriefings revealed that participants were unaware of whether they received tACS or sham stimulation. The order of trials and stimulation type (tACS, sham) was individually randomized.

In experiment 1 (Fig. 1B), click detection performance was assessed using a two-interval two-alternative forced-choice task. This task was chosen because it is less prone to induce fluctuations in perceivers' strategies (i.e., response bias) than more criterion-dependent yes/no tasks as used in Neuling et al. (2012). Two gray digits ("1" and "2") were shown continuously on a screen. Each digit was slightly enlarged for 1s to indicate one of two consecutive observation intervals. The click train was presented in one of these intervals (selected at random with equal probability), whereas the other interval contained only silence. Listeners judged which of the two intervals contained the clicks by pressing a button and then received visual feedback. The inter-trial interval was on average 1.5s, depending on which phase condition was presented on the next trial. On each trial from a given condition, the visual stimuli were timed differently (±118.1ms relative to the first click) so as to provide no valid cue of the exact timing of the clicks.

In experiment 2 (Fig. 1C), target detection was assessed using a yes/no task and a criterion-free performance measure within the framework of signal detection theory (Macmillan & Creelman, 1991). On each trial, the masker was presented continuously for 45s during which several targets (on average six), all associated with the same phase condition, were presented (for details, see section 2.2.1). Consecutive targets were separated by non-target intervals whose duration varied randomly between 2 and 7s, with the constraint that overall these intervals would occupy half of the trial time (thus targets occupied the other half). Trials began and ended with non-target intervals. Listeners were instructed to press and hold a key as soon and as long as they heard a repetitive tone. The inter-trial interval lasted on average 5s, depending on which phase condition was presented on the next trial.

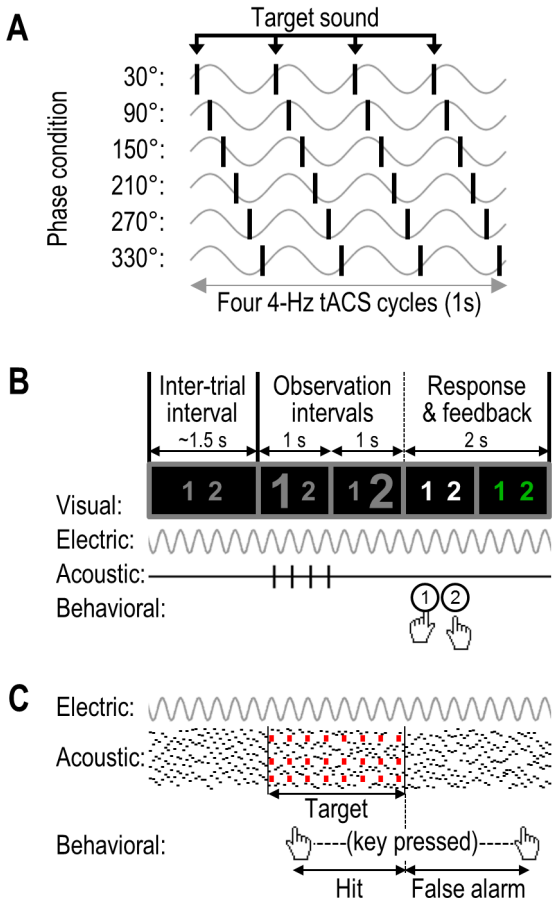


Figure 1. Experimental designs. Panel **A** illustrates the six tACS phase conditions for both experiments. Each tACS phase condition was characterized by a specific phase lag between the acoustic and electric stimuli. The gray wave schematizes four consecutive cycles of the 4-Hz tACS. The four black dashes sketch the target sound, which was either a train of four clicks presented at 4Hz (experiment 1) or a complex tone pulsed at 4Hz (experiment 2). Panel **B** exemplifies (from top to bottom) the visual, electric, and auditory stimulation, and a behavioral response from a trial in experiment 1. Listeners performed click detection in silence in a two-interval two-alternative forced-choice task. Panel **C** exemplifies a portion of a trial from experiment 2 (analogous to panel B, except that a spectrogram rather than sound wave is shown). Listeners were instructed to press and hold a key whenever they heard a repetitive tone target. The target was presented at unpredictable times and for unpredictable durations within an ongoing multi-tone masker.

2.3 Data analysis

For both experiments, data from the tACS runs were first concatenated and then a performance measure was extracted for each phase condition. For experiment 1, this measure was the accuracy, i.e., the number of correct responses divided by the number of trials (72 per phase condition). In experiment 2, the sensitivity index d' was computed based on the overall hit rate and overall false alarm rate, defined as the overall time during which the participant correctly or erroneously (respectively) reported the presence of the target, divided by half the overall trial time (3min corresponding to eight half trials, per phase condition).

For each of the extracted performance measures, a 'time series' (i.e., the six phase conditions spanning a 4-Hz period) was reconstructed. To compensate for possible inter-individual variations in 'best phase' (i.e., the phase for which performance was best), the maximum of the reconstructed series was aligned to the 90°-point and the series was phase-wrapped (Busch et al., 2009; Ng et al., 2012). Following this phase alignment—under the main hypothesis that the phase of the 4-Hz tACS modulates performance—phases 30°-150° should delimit a positive 4-Hz

half-cycle, whereas phases 210°-330° should delimit the opposite (negative) half-cycle. To verify this prediction, performance was averaged across the hypothesized positive half-cycle (phases 30° and 150°) and across the hypothesized negative half-cycle (phases 210°, 270°, and 330°), and then the two resulting averages were compared. To avoid circularity, the alignment point (90°) was excluded from this analysis. To test whether the hypothesized phase effect cycled specifically at 4Hz, spectral density was computed from the reconstructed series, and then the magnitudes of the individual frequency components were compared. Only the first three harmonics (4Hz, 8Hz, and 12Hz) could be resolved with this frequency analysis, due to the limited number of data points and sampling rate that could be obtained in the experiments (six points spanning a 4-Hz period).

Baseline was defined as the overall performance during the sham run, excluding the trials that were presented during the ramps at the beginning and end of the run.

2.4 Participants

Twenty-one paid volunteers (eight females, ages: 20-38 years) participated in the study so far (experiment 1: N=14, experiment 2: N=7). They reported no history of neurological, psychiatric, or hearing disorders, were eligible to undergo tACS as assessed by prior screening, and gave their written informed consent before taking part. They had normal hearing (defined as hearing thresholds of less than 25 dB HL at 250, 500, 750, 1000, 1500, 3000, and 4000Hz), except for one participant who had mild hearing loss in the left ear for the two highest frequencies. The experimental procedures were approved by the local ethics committee of the Faculty of Psychology and Neuroscience of Maastricht University (*Ethische Commissie Psychologie*).

3 Results

Figure 2A shows listeners' average performance in experiment 1 as a function of phase condition. Performance was above baseline in the 30° and 150° conditions, whereas it was below baseline in the 210°, 270°, and 330° conditions. Furthermore, performance was similar at 30° and 150°, and at 210° and 330° respectively, i.e., it was approximately symmetric with respect to the presumed best phase (90°). Performance at best phase was on average 7.8±4.3% better than at the opposite phase (270°) and 7.5±7.6% better than baseline (mean±sem across participants). Although on average the best performance and worst performance (210°, see Fig. 2A) were not associated with exactly opposite phases, possibly due to the contribution of higher harmonics, these observations match well the characteristics of the hypothesized 4-Hz cycle.

For Figure 2B, performance in experiment 1 was averaged across phase conditions to facilitate comparison of the hypothesized positive half-cycle (30°, 150°) with the hypothesized negative half-cycle (210°, 270°, 330°) and baseline. Second-level (random-effects) statistical group analyses using paired t-tests revealed an effect of half-cycle polarity on performance (positive vs. negative half-cycle: $t(13)=2.48$, corrected $p=0.042$); thus the tACS phase modulated listeners' auditory perception. Comparison with baseline revealed no effect (positive half-cycle vs. baseline: $t(13)=0.72$, corrected $p=0.37$; negative half-cycle vs. baseline: $t(13)=-0.32$, corrected $p=0.38$).

Figure 2C shows the magnitude spectrum of the reconstructed series, for experiment 1. Magnitude peaked at the tACS frequency component (4Hz) and decreased monotonically at the higher harmonics (8 and 12Hz). While the latter could explain a significant proportion of the overall variance in the series, this variance was explained best by a 4-Hz sinusoid; thus the phase-induced differences cycled predominantly at the tACS frequency. This observation was supported by a random-effects one-way ANOVA including frequency as factor, which revealed a main effect of frequency on spectral magnitude ($F(2,26)=3.93$, $p=0.032$). Post-hoc paired t-tests further showed that the 4-Hz component was significantly stronger than the 12-Hz component ($t(13)=2.65$, corrected $p=0.020$) but not the 8-Hz component ($t(13)=0.70$, corrected $p=0.25$).

Analogously, Figures 2D-2F show listeners' performance based on d', for experiment 2. These data are presented here only at the descriptive level due to the limited sample size (N=7). As for experiment 1, performance appears to be on average slightly better during the hypothesized positive 4-Hz half-cycle than during the opposite 4-Hz half-cycle or baseline (Figs. 2D, E). The magnitude spectrum shows more clearly that the overall variance in the reconstructed series seems to be explained better by a sinusoid at the tACS frequency than by sinusoids at higher harmonics (Fig. 2F). Statistical analyses based on a larger sample, which is being acquired at the moment, will assess the significance of these differences.

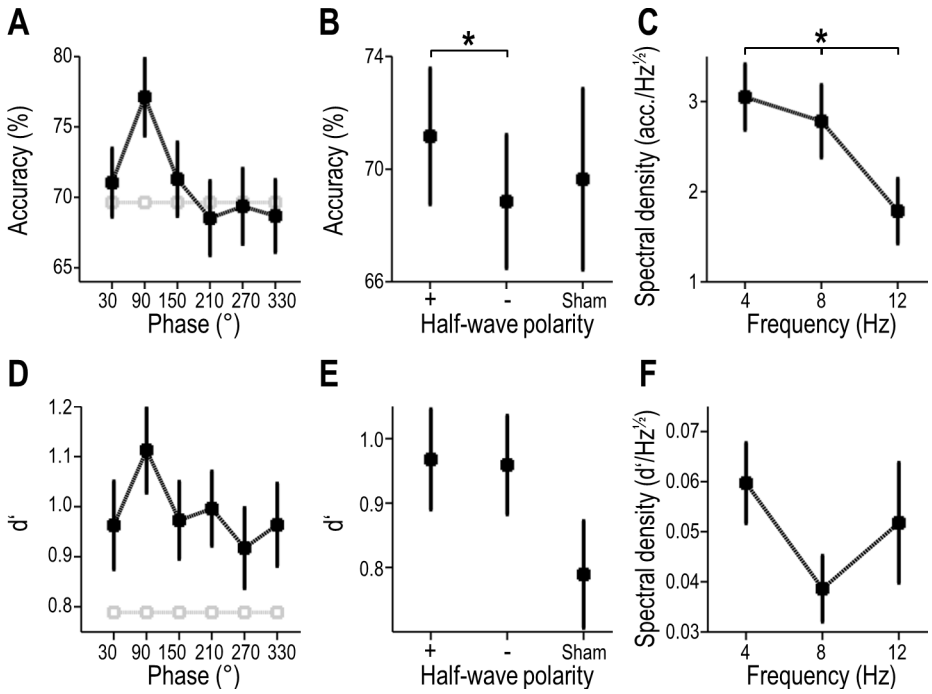


Figure 2. Results. The waveform in panel **A** shows behavioral performance (mean±sem across listeners) as a function of tACS phase condition (see Fig. 1A) for experiment 1. For reference, the gray horizontal line shows performance under sham stimulation. Panel **B** shows averages of these data, i.e., the hypothesized positive half-cycle (average of conditions 30° and 150°), the negative half-cycle (average of conditions 210°, 270°, and 330°), and sham stimulation. Panel **C** shows the magnitude spectrum of the performance waveform (mean±sem across listeners' individual magnitude spectra). Analogously, panels **D**, **E**, and **F** show preliminary data for experiment 2.

4 Discussion

The results from experiment 1 show that click detection under 4-Hz tACS is modulated by the tACS phase: performance was significantly better during one half of the 4-Hz tACS cycle than during the opposite half. Performance without tACS was intermediate, but did not differ significantly. The tACS phase effect was strongest at the tACS frequency but not strictly frequency-specific as it extended to higher harmonics. Overall, these data replicate the findings by Neuling et al. (2012), indicating that tACS phase effects on auditory perception are reproducible. They further extend these previous findings, which were based on single tone pips and single-channel alpha-tACS, to new stimuli including periodic clicks and dual-channel delta/theta-tACS. Considering that the clicks were presented at the tACS frequency and their audibility was found to depend on tACS phase, a possible interpretation is that tACS essentially enhanced the loudness of those click trains that were temporally coherent (rather than anti-coherent) with it. Such an 'envelope enhancement' would have interesting implications for

auditory scene analysis problems that depend on the relative salience of concurrent sounds and benefit from tracking the amplitude envelope of the target sound.

This idea is being addressed in experiment 2 using an informational masking paradigm, where an otherwise audible modulated target sound is rendered inaudible by spectrally non-overlapping concurrent distractor sounds. Based on magnetoencephalography findings, Elhilali et al. (2009) have proposed that the temporal coherence of cortical activity patterns with the modulated target sound plays a pivotal role for the perceptual segregation of the target from the distractors. Whether tACS phase modulates this temporal coherence, and therewith the audibility of the target, is still unclear. The preliminary data presented here show trends toward a possible effect of 4-Hz tACS phase, as in experiment 1.

Observations of actual benefits of tACS compared with no or sham stimulation are rather modest so far. In experiment 1, click detection at the best tACS phase was on average 7% more accurate than under sham stimulation. However, this difference did not reach statistical significance. Neuling et al. (2012) observed a 0.3-dB decrease in tone detection threshold compared with no tACS. A direct comparison of effect sizes across these tACS studies or the studies using acoustic entrainment stimuli is hampered by several differences including the different measures of interest. Future studies may identify which stimuli (single or dual channel tACS; electric or acoustic entrainment stimulus; 3, 4, 5, or 10Hz modulation frequency) are most effective. More generally, there is need for a systematic optimization of tACS parameters on an individual basis in order to maximize possible perceptual benefits of tACS. For example, the tACS frequency could be chosen to match the individual's dominant ongoing cortical frequency as identified beforehand with EEG (Zaehle et al., 2010; Ali et al., 2013). Similarly, tACS electrode positions could be chosen based on more realistic current flow simulations using as head model the individual's anatomy obtained beforehand with MRI (e.g., (Wagner et al., 2014)). Due to the limited spatial specificity of transcranial stimulation, individual tonotopic locations cannot be targeted as with multi-channel cochlear or brainstem implants. Nevertheless, the possibility to transmit acoustic envelope information non-invasively and bypass (deficits in) the auditory pathway up to the cortex might make transcranial stimulation a valuable tool.

Acknowledgments

I would like to thank Alexander Sack and Daniel Hauke for help with obtaining the data. This work was supported by Veni grant 451-11-014 from the Netherlands Organization for Scientific Research.

References

Ali, M. M., Sellers, K. K., & Frohlich, F. (2013). Transcranial alternating current stimulation modulates large-scale cortical network activity by network resonance. *The Journal of Neuroscience, 33*(27), 11262-11275.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences, 16*(7), 390-398.

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., Emerson, R. G., & Schroeder, C. E. (2011). Tuning of the human neocortex to the temporal dynamics of attended events. *The Journal of Neuroscience, 31*(9), 3176-3185.

Brignani, D., Ruzzoli, M., Mauri, P., & Miniussi, C. (2013). Is transcranial alternating current stimulation effective in modulating brain oscillations? *PLoS One, 8*(2), e56589.

Busch, N. A., Dubois, J., & VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *The Journal of Neuroscience, 29*(24), 7869-7876.

de Graaf, T. A., Gross, J., Paterson, G., Rusch, T., Sack, A. T., & Thut, G. (2013). Alpha-band rhythms in visual task performance: phase-locking by rhythmic sensory stimulation. *PLoS One, 8*(3), e60035.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *The Journal of Neuroscience, 33*(13), 5728-5735.

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology, 7*(6), e1000129.

Frohlich, F., & McCormick, D. A. (2010). Endogenous electric fields may guide neocortical network activity. *Neuron, 67*(1), 129-143.

Gomez-Ramirez, M., Kelly, S. P., Molholm, S., Sehatpour, P., Schwartz, T. H., & Foxe, J. J. (2011). Oscillatory sensory selection mechanisms during intersensory attention to rhythmic auditory and visual inputs: a human electrocorticographic investigation. *The Journal of Neuroscience, 31*(50), 18556-18567.

Gutschalk, A., Micheyl, C., & Oxenham, A. J. (2008). Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biology, 6*(6), e138.

Helfrich, R. F., Schneider, T. R., Rach, S., Trautmann-Lengsfeld, S. A., Engel, A. K., & Herrmann, C. S. (2014). Entrainment of brain oscillations by transcranial alternating current stimulation. *Current biology, 24*(3), 333-339.

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences of the United States of America, 109*(49), 20095-20100.

Henry, M. J., Herrmann, B., & Obleser, J. (2014). Entrained neural oscillations in multiple frequency bands comodulate behavior. *Proceedings of the National Academy of Sciences of the United States of America, 111*(41), 14935-14940.

Herrmann, C. S., Rach, S., Neuling, T., & Struber, D. (2013). Transcranial alternating current stimulation: a review of the underlying mechanisms and modulation of cognitive processes. *Frontiers in human neuroscience, 7*, 279.

Jaegle, A., & Ro, T. (2014). Direct control of visual perception with phase-specific modulation of posterior parietal cortex. *Journal of cognitive neuroscience, 26*(2), 422-432.

Jung, Y. J., Kim, J. H., & Im, C. H. (2013). COMETS: A MATLAB toolbox for simulating local electric fields generated by transcranial direct current stimulation (tDCS). *Biomedical Engineering Letters, 3*, 39-46.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience, 30*(2), 620-628.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology, 94*(3), 1904-1911.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron, 77*(4), 750-761.

Liegeois-Chauvel, C., Lorenzi, C., Trebuchon, A., Regis, J., & Chauvel, P. (2004). Temporal envelope processing in the human left and right auditory cortices. *Cerebral Cortex, 14*(7), 731-740.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: a user's guide*. Cambridge: Cambridge UP.

Marshall, L., Helgadottir, H., Molle, M., & Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature, 444*(7119), 610-613.

Neuling, T., Rach, S., & Herrmann, C. S. (2013). Orchestrating neuronal networks: sustained after-effects of transcranial alternating current stimulation depend upon brain states. *Frontiers in human neuroscience, 7*, 161.

Neuling, T., Rach, S., Wagner, S., Wolters, C. H., & Herrmann, C. S. (2012). Good vibrations: oscillatory phase shapes perception. *Neuroimage, 63*(2), 771-778.

Ng, B. S., Schroeder, T., & Kayser, C. (2012). A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *The Journal of Neuroscience, 32*(35), 12268-12276.

Nitsche, M. A., Cohen, L. G., Wassermann, E. M., Priori, A., Lang, N., Antal, A., Paulus, W., Hummel, F., Boggio, P. S., Fregni, F., & Pascual-Leone, A. (2008). Transcranial direct current stimulation: State of the art 2008. *Brain stimulation, 1*(3), 206-223.

Ozen, S., Sirota, A., Belluscio, M. A., Anastassiou, C. A., Stark, E., Koch, C., & Buzsaki, G. (2010). Transcranial electric stimulation entrains cortical neuronal populations in rats. *The Journal of Neuroscience, 30*(34), 11476-11485.

Paulus, W. (2011). Transcranial electrical stimulation (tES - tDCS; tRNS, tACS) methods. *Neuropsychological rehabilitation, 21*(5), 602-617.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex, 23*(6), 1378-1387.

Pogosyan, A., Gaynor, L. D., Eusebio, A., & Brown, P. (2009). Boosting cortical activity at Beta-band frequencies slows movement in humans. *Current Biology, 19*(19), 1637-1641.

Romei, V., Gross, J., & Thut, G. (2012). Sounds reset rhythms of visual cortex and corresponding human visual perception. *Current biology, 22*(9), 807-813.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in neurosciences, 32*(1), 9-18.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

Spaak, E., de Lange, F. P., & Jensen, O. (2014). Local entrainment of alpha oscillations by visual stimuli causes cyclic modulation of perception. *The Journal of Neuroscience, 34*(10), 3536-3544.

Stefanics, G., Hangya, B., Hernadi, I., Winkler, I., Lakatos, P., & Ulbert, I. (2010). Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *The Journal of Neuroscience, 30*(41), 13578-13585.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 31*(1), 164-171.

Wagner, S., Rampersad, S. M., Aydin, U., Vorwerk, J., Oostendorp, T. F., Neuling, T., Herrmann, C. S., Stegeman, D. F., & Wolters, C. H. (2014). Investigation of tDCS volume conduction effects in a highly realistic head model. *Journal of neural engineering, 11*(1), 016002.

Zaehle, T., Rach, S., & Herrmann, C. S. (2010). Transcranial alternating current stimulation enhances individual alpha activity in human EEG. *PLoS One, 5*(11), e13766.

Zaghi, S., Acar, M., Hultgren, B., Boggio, P. S., & Fregni, F. (2010). Noninvasive brain stimulation with low-intensity electrical currents: putative mechanisms of action for direct and alternating current stimulation. *The Neuroscientist, 16*(3), 285-307.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*(5), 980-991.

Zoefel, B., & Heil, P. (2013). Detection of Near-Threshold Sounds is Independent of EEG Phase in Common Frequency Bands. *Frontiers in psychology, 4*, 262.

Chapter 40

Functional Organization of the Ventral Auditory Pathway

Yale E. Cohen[1,2,3], Sharath Bennur[1], Kate Christison-Lagay[4], Adam M. Gifford[4], Joji Tsunada[1]

[1] *Department of Otorhinolaryngology, University of Pennsylvania, Philadelphia, USA*
[2] *Department of Neuroscience, University of Pennsylvania, Philadelphia, USA*
[3] *Department of Bioengineering, University of Pennsylvania, Philadelphia, USA*
[4] *Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, USA*

Contact author: YEC, ycohen@mail.med.upenn.edu

Abstract

The fundamental problem in audition is determining the mechanisms required by the brain to transform an unlabelled mixture of auditory stimuli into coherent perceptual representations. This process is called auditory-scene analysis. The perceptual representations that result from auditory-scene analysis are formed through a complex interaction of perceptual grouping, attention, categorization and decision-making. Despite a great deal of scientific energy devoted to understanding these aspects of hearing, we still do not understand (1) how sound perception arises from neural activity and (2) the causal relationship between neural activity and sound perception. Here, we review the role of the "ventral" auditory pathway in sound perception. We hypothesize that, in the early parts of the auditory cortex, neural activity reflects the auditory properties of a stimulus. However, in latter parts of the auditory cortex, neurons encode the sensory evidence that forms an auditory decision and are causally involved in the decision process. Finally, in the prefrontal cortex, which receives input from the auditory cortex, neural activity reflects the actual perceptual decision. Together, these studies indicate that the ventral pathway contains hierarchical circuits that are specialized for auditory perception and scene analysis.

Keywords: auditory perception, decision-making, categorization, perceptual grouping, ventral auditory pathway

1 Introduction

Hearing and communication present a variety of challenges to the auditory system. To be heard and understood, the auditory brain must transform a time-varying acoustic stimulus into a perceptual representation; that is, a sound.

Auditory perception is associated with a number of computational processes, which may act in parallel or serial, including: perceptual grouping, decision-making, attention, and categorization. (1) Perceptual grouping is a form of feature-based stimulus segmentation that determines whether acoustic events will be grouped into a single sound or be segregated into distinct sounds (Bregman, 1990). (2) Auditory decision-making is a computational process in which the brain interprets sensory information in order to detect, discriminate, or identify the source or content of auditory stimuli (Gold & Shadlen, 2007). Did I hear the stimulus? From where and whom did it come? What does it tell me? How can I use this information to plan an action? (3)

Although attention is not always necessary, our awareness of a sound can be influenced by attention (Alain & Arnott, 2000; Micheyl et al., 2003; Fritz et al., 2005; Shinn-Cunningham, 2008; Snyder et al., 2012; Gutschalk et al., 2015). For example, we can choose whether to listen to—or ignore—the first violin, the string section, or even the whole orchestra. Likewise, we can selectively attend to the particular features in a person's voice that allow a listener to identify the speaker. (4) In auditory categorization, sounds are classified based on their acoustic features or more processed forms of information (e.g., semantic knowledge), providing an efficient means to interact with stimuli in our environment (Ashby & Berretty, 1997; Gifford et al., 2014). For example, when we hear the word "Groningen" from different speakers, we can categorize the gender of each speaker based on the pitch of the speaker's voice. On the other hand, to analyze the linguistic content transmitted by a speech sound, we can ignore the unique pitch, timbre etc. of each speaker and categorize the sound into the distinct word category "Groningen".

It is thought that the neural computations and processes that mediate auditory perception are found in the ventral auditory pathway (Rauschecker & Scott, 2009; Romanski & Averbeck, 2009; Hackett, 2011; Bizley & Cohen, 2013). In rhesus monkeys, this pathway begins in core auditory cortex—specifically, primary auditory cortex (A1) and the rostral field. These core areas project to the middle lateral (ML) and anterolateral belt (AL) regions of auditory cortex. In turn, these belt regions project directly and indirectly to the ventrolateral prefrontal cortex (vlPFC).

It is important to briefly comment on the contribution of the dorsal ("spatial") pathway to auditory perception (Rauschecker, 2012; Cloutman, 2013). Spatial information can act as a grouping cue to assist the segregation of an acoustic stimulus into discrete sounds. For example, when a rhythmic sequence of identical sound bursts is presented from a single location, it is often perceived as one source. But, when the sound sequence is presented from two different locations, it can be perceived as two sounds (Middlebrooks & Onsan, 2012; Middlebrooks & Bremen, 2013). Such findings suggest that a mixture of spatial and non-spatial auditory cues from both the dorsal and ventral pathways may be needed in order to create a coherent auditory-perceptual representation that guides behavior.

Nonetheless, in this review, we focus on the hierarchical processing that occurs at different stages in the ventral auditory pathway. In particular, we identify—or, at least, suggest—the unique contributions of these different processing stages to auditory perception and categorization, with an acknowledgment that associating any single brain region with a particular computation oversimplifies the complexity of the auditory brain. Indeed, it is well known that neurons become increasingly sensitive to more complex stimuli along the beginning stages of the ventral auditory pathway (e.g., between the core and belt regions of the auditory cortex). For example, core neurons are more sharply tuned for tone bursts than neurons in the lateral belt, whereas lateral-belt neurons are more selective for particular spectrotemporal features of complex sounds, such as vocalizations (Rauschecker & Tian, 2000). Here, though, we review hierarchical processing in the ventral pathway by focusing on those studies in which neural activity was collected simultaneously while a listener was engaged in an auditory task.

2 Neural correlates of auditory perception along the ventral auditory pathway

A1's role in auditory perception is controversial. Part of that controversy stems from the putative role of A1 in processing auditory "objects" (Nelken, 2008). We will take the position that auditory objects are analogous to perceptual representations (i.e., sounds) (Bizley & Cohen, 2013). As a consequence of this definition, if a neuron encodes an auditory object, it should be modulated by a listener's perceptual reports. That is, by holding a stimulus constant and testing whether neural activity is modulated by a listener's reports, neural activity that is associated with the acoustic features of the stimulus can be dissociated from neural activity associated with

the perceptual report. Thus, neurons with complex tuning properties or even those modulated by components of a task (Brosch et al., 2005; Fritz et al., 2005) may contribute to the construction of a perceptual representation; but by themselves do not offer direct evidence of a perceptual representation.

There has been a recent set of literature implicating A1 in auditory perceptual decision-making (Riecke et al., 2009; Kilian-Hutten et al., 2011; Niwa et al., 2012; Riecke et al., 2012; Bizley et al., 2013). In one study, ferrets were asked to report changes in a sound's pitch, and it was found that both local-field potentials and spiking activity in A1 were modulated by the ferrets' pitch judgments. Niwa and colleagues have also shown that A1 single-unit activity is modulated by monkeys' reports during a task in which monkeys reported whether or not a sound was amplitude modulated. Finally, human-imaging studies have revealed that regions of core auditory cortex are modulated by listener's reports of the identity of an ambiguous speech sound.

However, a different body of work suggests that A1 does not encode auditory decisions. For example, when monkeys discriminate between two types of acoustic flutter, A1 activity is not modulated by the monkeys' choices (Lemus et al., 2009b). Preliminary studies from our laboratory also support the hypothesis that A1 activity does not reflect choice. In one study, we trained monkeys to listen to an auditory stimulus and report whether they heard one or two auditory streams (Bregman, 1990; Christison-Lagay & Cohen, 2014). The auditory stimulus consisted of an alternating sequence of tone bursts at two different frequencies (frequency A and B). Critically, at intermediate frequency differences (e.g., 3 or 5 semitones), listeners hear one or two auditory streams on alternate trials. To date, our data suggest that A1 activity—particularly, on trials with intermediate frequency differences—does not reflect the monkeys' reports but, instead, simply reflects the acoustic features of the tone-burst sequence. In a second set of studies, we asked monkeys to detect a target sound that was embedded in a noisy background, while simultaneously recording A1 activity. Once again, our preliminary data indicate that, whereas the frequency and sound level of the target modulated A1 activity, the monkeys' reports of target detection did not.

What could be the bases for these apparent divergent sets of findings? We posit that these differences can be attributed to several non-exclusive possibilities. One possibility may due to the relative perceptual and cognitive demands of the behavioural task: tasks with different demands might differentially engage neurons in core auditory cortex (Bizley & Cohen, 2013; Nienborg & Cumming, 2014). A second possibility focuses on how choice-related activity itself is analyzed. In choice analyses, it is imperative to restrict the analysis to those trials in which neural modulation related to choice can be clearly disassociated from the auditory stimulus. If this analysis is not carefully conducted, apparent choice activity may be confabulated with stimulus-related activity. Finally, choice-related activity may not reflect a casual contribution of the auditory cortex to decision-making but may simply reflect feedback from higher choice-sensitive areas (Nienborg & Cumming, 2009) or the structure of the correlated noise (Nienborg et al., 2012).

In the belt regions (ML and AL) of the ventral pathway, several lines of study from our laboratory suggest that this is not the case; but also see (Niwa et al., 2013). While monkeys categorized speech sounds, we tested whether neural activity was modulated by monkeys' categorical judgements. We found that AL neurons were not modulated by the monkeys' reports (Tsunada et al., 2011; Tsunada et al., 2012). In a separate line of studies, we asked monkeys to listen to a sequence of tone bursts and report whether the sequence had a "low" or "high" pitch. Our preliminary findings suggest that neither ML nor AL activity was modulated by the monkeys' choices.

Do neurons in the different belt regions contribute differentially to auditory perception? The preliminary findings from our low-high pitch study suggest that AL neurons, but not ML neurons, might represent the sensory information (evidence) used to inform the monkeys'

perceptual decisions. Specifically, AL neurometric sensitivity appeared to be correlated with both psychometric sensitivity and the monkeys' choices. Consistent with these findings, AL may play a causal role in these auditory judgments: microstimulation of an AL site tends to shift the monkeys' reports toward the pitch associated with the site's frequency tuning.

Whereas a single cortical locus of decision-making has yet to materialize, decision-related activity is seen throughout the frontal lobe. vlPFC neurons are strongly modulated by monkeys' choices (Russ et al., 2008; Lee et al., 2009). Neural activity in the inferior frontal lobe of the human cortex is also modulated by choice when listeners judge the content of ambiguous speech sounds (Binder et al., 2004). Neural correlates relating to a listeners' decision on auditory-flutter stimuli have also been observed in the ventral premotor cortex (Lemus et al., 2009a). Interestingly, as noted above, the dorsal pathway also contributes to auditory perception; consistent with that notion, activity in the human parietal lobe is modulated by listeners' choices (Cusack, 2005).

In summary, we propose a model in which auditory information is hierarchically organized and processed in the ventral pathway. In early parts of the auditory cortex, neural activity encodes the acoustic features of an auditory stimulus and become increasingly sensitive to complex spectrotemporal properties (Rauschecker & Tian, 2000). In later regions of the auditory cortex, this information informs perceptual judgments. However, neural activity that reflects a listener's perceptual judgments does not become apparent until the frontal lobe.

3 Category representation in the ventral auditory pathway

Next, we review the manner in which auditory-category information is hierarchically organized in the ventral auditory pathway. In brief, we will highlight how feature-based categories are represented early; whereas in later parts of the pathway, we find representations of "abstract" categories, which combine acoustic information with mnemonic, emotional, and other information sources.

In core auditory cortex, neural activity codes the category membership of simple feature conjunctions. For example, categorical representations of frequency-contours have been identified (Ohl et al., 2001; Selezneva et al., 2006). These neurons encode the direction of a frequency contour (increasing or decreasing), independent of its specific frequency content. These categories may be found in the firing rates of individual neurons or may be observed as a result of population-level computations.

Categories for more complex stimuli, such as speech sounds and vocalizations, can be found in the lateral belt (Chang et al., 2010; Steinschneider et al., 2011; Tsunada et al., 2011; Steinschneider, 2013). For example, AL neurons respond categorically, and in a manner consistent with listeners' behavioral reports, to morphed versions of two speech sounds ("bad" and "dad"). AL neurons also respond categorically to species-specific vocalizations; however, the degree to which AL (and listeners) can categorize these vocalizations is constrained by the vocalizations' acoustic variability (Christison-Lagay et al., 2014). In humans, the superior temporal gyrus is categorically and hierarchically organized by speech sounds (Binder et al., 2000; Chang et al., 2010; Leaver & Rauschecker, 2010): phoneme categories are found in the middle aspect; word categories in the anterior-superior aspect; and phrases in the most anterior aspect (DeWitt & Rauschecker, 2012; Rauschecker, 2012).

Beyond the auditory cortex, neurons represent categories that are formed based on the abstract information that is transmitted by sounds. vlPFC neurons represent the valence of food-related calls (e.g., high quality food vs. low quality food) (Gifford et al., 2005). That is, vlPFC neurons encode the "referential" information that is transmitted by vocalizations, independent of differences in their acoustic properties. Prefrontal activity also contributes to the formation of categories that reflect the emotional valence of a speaker's voice (Fecteau et al., 2005) as well as the semantic information transmitted by multisensory stimuli (Joassin et al., 2011; Werner &

Noppeney, 2011; Hu et al., 2012). Together, these studies are consistent with the idea that ventral auditory pathway is an information-processing pathway that more complex stimuli and categories are processed in a hierarchically organized manner.

4 Future questions

Of course, several fundamental questions remain. First, as alluded to above, understanding how feedforward versus feedback information contributes to neural correlates of perceptual judgments remains an open question. Second, the degree to which different types of auditory judgements differentially engage different components of the ventral pathway has yet to be fully articulated. A third question is to identify how the different computational processes (e.g., perceptual grouping, attention, decision-making, and categorization) that underlie auditory perception interact with one another. For example, it remains an open issue as to whether and how attention differentially modulates neural correlates of auditory perception at different hierarchical levels of the ventral auditory pathway (e.g., A1 versus AL) (Atiani et al., 2014). Another example is to identify the potential interactions between auditory perceptual grouping and decision-making. Finally, it is important to identify the manner by which the dorsal and ventral auditory pathways interact in order to form a consistent and coherent representation of the auditory scene.

References

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. Frontiers in Biosci, 5, d202-212.

Ashby, F. G., & Berretty, P. M. (1997). Categorization as a special case of decision-making or choice. In A. A. J. Marley (Ed.), Choice, decision, and measurement: Essays in honor of R. Duncan Luce (pp. 367-388). Hillsdale, NH: Lawrence Erlbaum Associates.

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. Neuron, 82(2), 486-499. doi: 10.1016/j.neuron.2014.02.029

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. Nat Neurosci, 7, 295-301.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex, 10(5), 512-528.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where, and how of auditory-object perception. Nat Rev Neurosci, 14, 693-707.

Bizley, J. K., Walker, K. M., Nodal, F. R., King, A. J., & Schnupp, J. W. (2013). Auditory cortex represents both pitch judgments and the corresponding acoustic cues. [Research Support, Non-U.S. Gov't]. Curr Biol, 23(7), 620-625.

Bregman, A. S. (1990). Auditory Scene Analysis. Boston, MA: MIT Press.

Brosch, M., Selezneva, E., & Scheich, H. (2005). Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. J Neurosci, 25(29), 6797-6806.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. Nat Neurosci, 13, 1428-1432.

Christison-Lagay, K. L., & Cohen, Y. E. (2014). Behavioral correlates of auditory streaming in rhesus macaques. Hear Res, 309, 17-25.

Christison-Lagay, K. L., Bennur, S., Lee, J. H., Blackwell, J., Schröder, T., & Cohen, Y. E. (2014). Natural variability in species-specific vocalizations constrains behavior and neural activity. Hear Res, 312, 128-142.

Cloutman, L. L. (2013). Interaction between dorsal and ventral processing streams: where, when and how? Brain Lang, 127(2), 251-263. doi: 10.1016/j.bandl.2012.08.003

Cusack, R. (2005). The Intraparietal Sulcus and Perceptual Organization. J Cogn Neurosci, 17, 641-651.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A, 109(8), E505-514.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2005). Sensitivity to voice in human prefrontal cortex. J Neurophysiol, 94(3), 2251-2254. doi: 10.1152/jn.00329.2005

Fritz, J., Elhilali, M., & Shamma, S. (2005). Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. Hear Res, 206(1-2), 159-176. doi: S0378-5955(05)00083-3 [pii]

10.1016/j.heares.2005.01.015

Gifford, A. M., Cohen, Y. E., & Stocker, A. A. (2014). Characterizing the impact of category uncertainty on human auditory categorization behavior. PloS Comp Biol, 10, e1003715.

Gifford, G. W., 3rd, MacLean, K. A., Hauser, M. D., & Cohen, Y. E. (2005). The neurophysiology of functionally meaningful categories: macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. Journal of Cognitive Neuroscience, 17(9), 1471-1482.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annu Rev Neurosci, 30, 535-574.

Gutschalk, A., Rupp, A., & Dykstra, A. R. (2015). Interaction of Streaming and Attention in Human Auditory Cortex. PLoS One, 10(3).

Hackett, T. A. (2011). Information flow in the auditory cortical network. Hear Res, 271, 133-146.

Hu, Z., Zhang, R., Zhang, Q., Liu, Q., & Li, H. (2012). Neural correlates of audiovisual integration of semantic category information. Brain Lang, 121(1), 70-75. doi: 10.1016/j.bandl.2012.01.002

Joassin, F., Maurage, P., & Campanella, S. (2011). The neural network sustaining the crossmodal processing of human gender from faces and voices: an fMRI study. Neuroimage, 54(2), 1654-1661. doi: 10.1016/j.neuroimage.2010.08.073

Kilian-Hutten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. J Neurosci, 31(5), 1715-1720. doi: 10.1523/jneurosci.4572-10.2011

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci, 30, 7604-7612.

Lee, J. H., Russ, B. E., Orr, L. E., & Cohen, Y. E. (2009). Prefrontal activity predicts monkeys' decisions during an auditory category task. Fron Integrative Neuroscience, 3(16).

Lemus, L., Hernandez, A., & Romo, R. (2009a). Neural encoding of auditory discrimination in ventral premotor cortex. Proc Natl Acad Sci U S A, 106(34), 14640-14645. doi: 10.1073/pnas.0907505106

Lemus, L., Hernandez, A., & Romo, R. (2009b). Neural codes for perceptual discrimination of acoustic flutter in the primate auditory cortex. Proc Natl Acad Sci U S A, 106, 9471-9476.

Micheyl, C., Carlyon, R. P., Shtyrov, Y., Hauk, O., Dodson, T., & Pullvermuller, F. (2003). The neurophysiological basis of the auditory continuity illusion: a mismatch negativity study. J Cogn Neurosci, 15(5), 747-758. doi: 10.1162/089892903322307456

Middlebrooks, J. C., & Onsan, Z. A. (2012). Stream segregation with high spatial acuity. J Acoust Soc Am, 132, 3896-3911.

Middlebrooks, J. C., & Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. J Neurosci, 33, 10986-11001.

Nelken, I. (2008). Processing of complex sounds in the auditory system. Curr Opin Neurobiol, 18, 413-417.

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature, 459(7243), 89-92. doi: 10.1038/nature07821

Nienborg, H., & Cumming, B. G. (2014). Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. J Neurosci, 34(10), 3579-3585. doi: 10.1523/jneurosci.2340-13.2014

Nienborg, H., Cohen, M. R., & Cumming, B. G. (2012). Decision-related activity in sensory neurons: correlations among neurons and with behavior. [Review]. Annual Review of Neuroscience, 35, 463-483.

Niwa, M., Johnson, J. S., O'Connor, K. N., & Sutter, M. L. (2012). Active engagement improves primary auditory cortical neurons' ability to discriminate temporal modulation. J Neurosci, 32(27), 9323-9334.

Niwa, M., Johnson, J. S., O'Connor, K. N., & Sutter, M. L. (2013). Differences between Primary Auditory Cortex and Auditory Belt Related to Encoding and Choice for AM Sounds. J Neurosci, 33(19), 8378-8395.

Ohl, F. W., Scheich, H., & Freeman, W. J. (2001). Change in pattern of ongoing cortical activity with auditory category learning. Nature, 412(6848), 733-736.

Rauschecker, J. P. (2012). Ventral and dorsal streams in the evolution of speech and language. Front Evol Neurosci, 4(7).

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci U S A, 97(22), 11800-11806.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci, 12(6), 718-724.

Riecke, L., Mendelsohn, D., Schreiner, C., & Formisano, E. (2009). The continuity illusion adapts to the auditory scene. Hear Res, 247(1), 71-77. doi: 10.1016/j.heares.2008.10.006

Riecke, L., Vanbussel, M., Hausfeld, L., Baskent, D., Formisano, E., & Esposito, F. (2012). Hearing an illusory vowel in noise: suppression of auditory cortical activity. J Neurosci, 32(23), 8024-8034. doi: 10.1523/jneurosci.0440-12.2012

Romanski, L. M., & Averbeck, B. B. (2009). The Primate Cortical Auditory System and Neural Representation of Conspecific Vocalizations. Annu Rev Neurosci, 32, 315-346.

Russ, B. E., Orr, L. E., & Cohen, Y. E. (2008). Prefrontal neurons predict choices during an auditory same-different task. Curr Biol, 18(19), 1483-1488.

Selezneva, E., Scheich, H., & Brosch, M. (2006). Dual Time Scales for Categorical Decision Making in Auditory Cortex. Curr Biol, 16(24), 2428-2433.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. Trends Cogn Sci, 12, 182-186.

Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, Awareness, and the Perception of Auditory Scenes. Front Psychol, 3.

Steinschneider, M. (2013). Phonemic representations and categories. In Y. E. Cohen, A. N. Popper & R. R. Fay (Eds.), Neural correlates of auditory cognition (Vol. 45, pp. 151-161): Springer-Verlag.

Steinschneider, M., Nourski, K. V., Kawasaki, H., Oya, H., Brugge, J. F., & Howard, M. A. (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. Cereb Cortex, 10, 2332-2347.

Tsunada, J., Lee, J. H., & Cohen, Y. E. (2011). Representation of speech categories in the primate auditory cortex. J Neurophysiol, 105, 2634-2646.

Tsunada, J., Lee, J. H., & Cohen, Y. E. (2012). Differential representation of auditory categories between cell classes in primate auditory cortex. J Physiol, 590, 3129-3139.

Werner, S., & Noppeney, U. (2011). The contributions of transient and sustained response codes to audiovisual integration. Cereb Cortex, 21(4), 920-931. doi: 10.1093/cercor/bhq161

Chapter 41

Neural Segregation of Concurrent Speech: Effects of Background Noise and Reverberation on Auditory Scene Analysis in the Ventral Cochlear Nucleus

Mark Sayles[1,2], Arkadiusz Stasiak[1], Ian M. Winter[1,*]

[1] *Centre for the Neural Basis of Hearing, The Physiological Laboratory, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK*
[2] *Laboratory of Auditory Neurophysiology, Campus Gasthuisberg, O&N II, Leuven, Belgium*

Contact author: IMW, imw1001@cam.ac.uk

Abstract

Concurrent complex sounds (e.g., two voices speaking at once) are perceptually disentangled into separate "auditory objects". This neural processing often occurs in the presence of acoustic-signal distortions from noise and reverberation (e.g., in a busy restaurant). A difference in periodicity between sounds is a strong segregation cue under quiet, anechoic conditions. However, noise and reverberation exert differential effects on speech intelligibility under "cocktail-party" listening conditions. Previous neurophysiological studies have concentrated on understanding auditory scene analysis under ideal listening conditions. Here, we examine the effects of noise and reverberation on periodicity-based neural segregation of concurrent vowels /a/ and /i/, in the responses of single units in the guinea-pig ventral cochlear nucleus (VCN): the first processing station of the auditory brain stem. In line with human psychoacoustic data, we find reverberation significantly impairs segregation when vowels have an intonated pitch contour, but not when they are spoken on a monotone. In contrast, noise impairs segregation independent of intonation pattern. These results are informative for models of speech processing under ecologically valid listening conditions, where noise and reverberation abound.

Keywords: vowels, double vowels, intonation, reverberation, background noise, auditory scene analysis, concurrent speech, perceptual segregation, inter-spike intervals, temporal envelope, brain stem, cochlear nucleus.

1 Introduction

A difference in periodicity between simultaneous complex sounds is a strong segregation cue under quiet, anechoic conditions (Brokx & Nooteboom, 1982). However, noise and reverberation can both degrade speech intelligibility under realistic "cocktail-party" listening conditions (Nabelek, 1993; Culling et al., 1994; Culling et al., 2003). Neurophysiological studies of concurrent-sound segregation have concentrated harmonic complex sounds, with a fundamental frequency (F0) difference ΔF0, heard under idealized (quiet, anechoic) conditions (e.g., Palmer, 1990; Keilson et al., 1997; Larsen et al., 2008). We examine the effects of noise and reverberation, separately, on periodicity-based neural segregation of ΔF0 concurrent vowels,

with and without simulated intonation, in the ventral cochlear nucleus (VCN); the first processing station in the auditory brain stem.

2 Methods

2.1 Animal Model

Experiments were carried out in accordance with the United Kingdom Animals (Scientific Procedures) Act (1986), with approval of the University of Cambridge Animal Welfare Ethical Review Board. Details of our recording techniques are available elsewhere (Sayles & Winter, 2008; Sayles et al., 2015). Adult guinea pigs (*Cavia porcellus*) were anesthetized with urethane and hypnorm (fentanyl/fluanisone). The cochlear nucleus was exposed via a posterior-fossa craniotomy and unilateral cerebellotomy. A glass-insulated tungsten microelectrode was advanced in the sagittal plane through the VCN, using a hydraulic microdrive. Upon isolation of a single unit, best frequency (BF) and threshold were determined. Units were classified on their responses to BF-tones.

2.2 Complex Stimuli

Stimuli were synthetic vowels /a/ and /i/, generated using a MATLAB implementation of the Klatt formant synthesizer. Formant frequencies were {0.7, 1.0, 2.4, 3.3} kHz for /a/, and {0.3, 2.2, 3.0, 3.7} kHz for /i/. Stimuli were presented monaurally; either alone, or as "double vowels" /a,i/ (Fig. 1). F0 was either static (125 Hz, or 250 Hz for /a/, and 100 Hz, or 200 Hz for /i/), or sinusoidally modulated at 5 Hz, by ± 2 semitones. Reverberation was added by convolution with real-room impulse responses recorded in a long corridor, at source-receiver distances of 0.32, 2.5, and 10 m (Tony Watkins; University of Reading, UK). We refer to these as "mild", "moderate" and "strong" reverberation. For noise-masked vowels, Gaussian noise (5-kHz bandwidth) was added at signal-to-noise ratios of {10, 3, 0} dB.

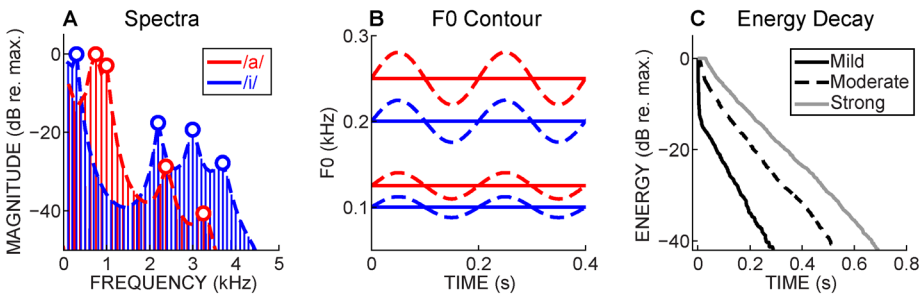


Figure 1. Synthetic-vowel stimuli. (A), Magnitude spectra. (B), Fundamental-frequency contours: solid lines, static-F0 vowels; dashed lines, modulated-F0 vowels. (C), Impulse-response energy-decay curves.

2.3 Analyses

Correlograms

Analyses are based on normalized shuffled inter-spike-interval distributions derived from the VCN-unit spike trains (Fig. 2; Joris et al., 2006; Sayles et al., 2015). We computed the *across*-spike-train shuffled inter-spike-interval distribution in a short (30-ms duration) rectangular time window, slid in 5-ms steps through the 400-ms duration response. We refer to these time-varying inter-spike-interval distributions as *correlograms*.

Periodic Templates

We applied a "periodicity-sieve" analysis to the correlograms to estimate the dominant period(s) in the inter-spike-interval statistics (e.g., Larsen et al., 2008); yielding *template-contrast functions*. To assess the statistical significance of peaks in the template-contrast function we used a bootstrap technique (permutation analysis; 1,000 replications: $p<0.01$ considered significant). Based on the template-contrast functions in response to double vowels, we computed the "periodicity-tagged" firing rate for each vowel of the mixture (similar to that proposed by Keilson et al., 1997); e.g., for the /a/ component:

$$\bar{R}_{/a/} = max\left(\left(\frac{\bar{R} \cdot C_{/a/}}{C_{/a/} + C_{/i/}}\right) - (0.5 \cdot \bar{R}),\ 0\right)$$

where $\bar{R}$ is the mean firing rate of that single unit to the double vowel /a,i/, and $C_{/a/}$ and $C_{/i/}$ are template-contrast values for the two double-vowel components, respectively.

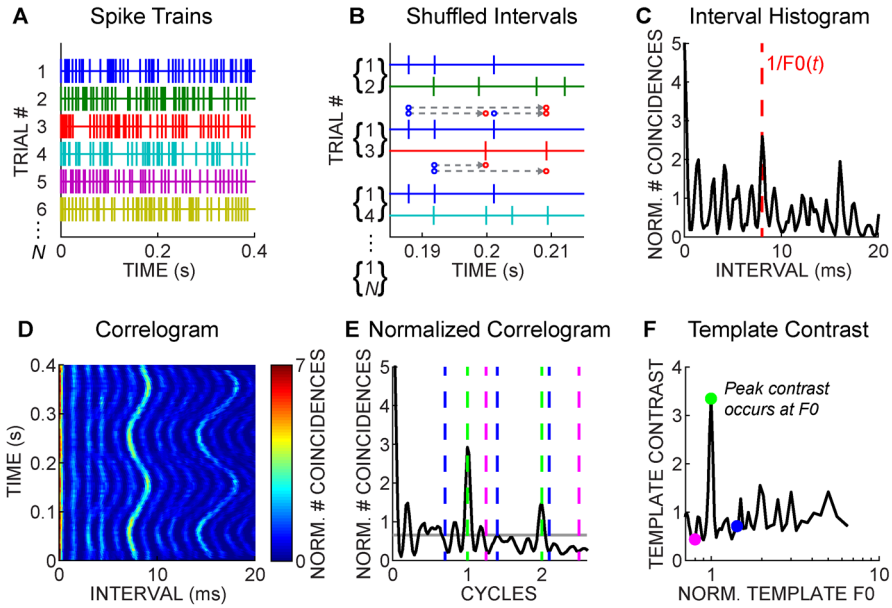


Figure 2. Example analyses for a single unit (Primary-like, BF = 2.1 kHz), responding to a single vowel /a/ with F0 modulated at 5 Hz (±2 semitones) around a mean F0 of 125 Hz. (A), Spike times were collected in response to *N* (typically, 50) repetitions of each stimulus. (B), Forward inter-spike intervals were calculated between all non-identical spike-train pairs, in 30-ms time windows. (C), Intervals were tallied in a histogram, and the analysis window slid in 5-ms steps to give the interval distribution as a function of time (D). (E), Time-varying interval distributions were normalized for instantaneous stimulus F0, and averaged over time. (F), Harmonic periodicity sieves were applied to compute the template-contrast function.

3 Results

We recorded responses to single and double vowels, in anechoic and reverberant conditions, from 129 units with BFs between 0.1 and 6 kHz (36 primary-like [PL/PN], 47 chopper [CT/CS], 24 onset [OC/OL], 19 low-frequency [LF], 3 unusual [UN]). From 52 of these, we also recorded responses to vowels in noise. The effects of noise and reverberation on double-vowel segregation are not equivalent. The results can be summarized as: (1), There is a strong interaction between F0 modulation (simulating intonated speech) and reverberation to reduce template contrast. (2), There is no interaction between F0 modulation and signal-to-noise ratio.

(3), Noise-induced deficits in neural periodicity-based double-vowel segregation are strongly BF-dependent, due to more total (masking) noise power passed by higher-BF filters. (4), Reverberation impairs neural segregation of intonated double vowels independent of BF, but has only marginally detrimental effects on segregation of double vowels with steady F0s.

Figure 3 shows the template contrast calculated from responses to the single vowel /a/ under ideal (anechoic, quiet) conditions, in the presence of "strong" reverberation, and in the presence of background noise at 0-dB SNR. There is a small reduction in template contrast due to reverberation alone (Fig. 3A and B), and a similar small reduction due to F0 modulation alone (Fig. 3A and D). However, the major detrimental effect on template contrast is an interaction between reverberation and F0 modulation (Fig. 3A and E). In the combined presence of reverberation and F0 modulation many units do not have a significant representation of the periodicity corresponding to this single vowel in their inter-spike-interval distributions; the many non-significant gray symbols in Fig. 3E. In contrast, the presence of masking noise reduces template contrast at the F0 period independent of F0 modulation, but in a BF-dependent manner with higher-BF units affected most strongly (Fig. 3C and F).

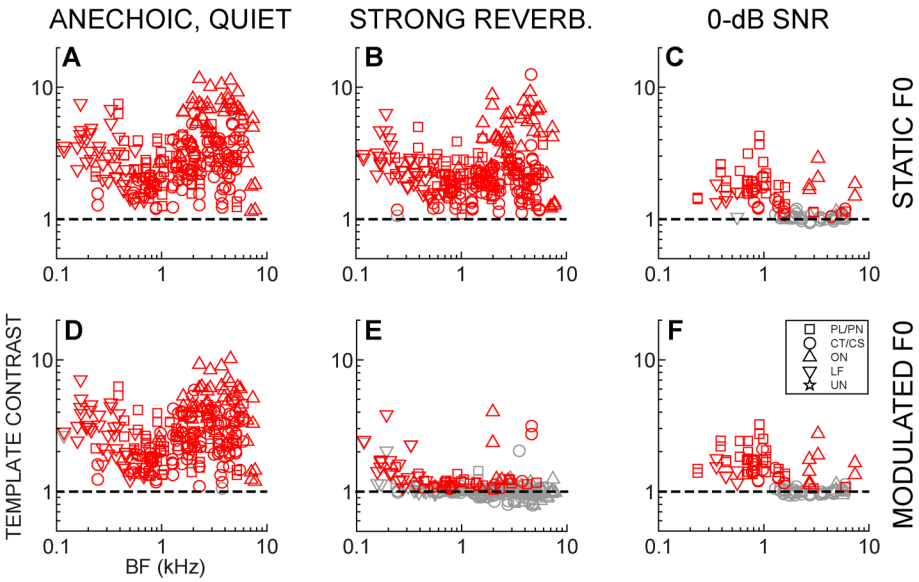


Figure 3. Template contrast *vs.* unit BF for responses to the single vowel /a/. *Top row*, responses to static-F0 vowels; *Bottom row*, responses to modulated-F0 vowels. Different symbols indicate unit type (legend in F). Red symbols indicate significant values (p<0.01), gray symbols are non-significant. Dashed line indicates contrast of 1, the expected value for a flat correlogram.

The interaction between reverberation and F0 modulation is further illustrated by plotting template contrast in response to modulated-F0 vowels against that in response to static-F0 vowels (Fig. 4). For both single- and double-vowel responses, the reduction in template contrast in the presence of reverberation is much greater in the modulated-F0 condition relative to the static-F0 condition (Fig. 4B and E). In the presence of noise, the maximum template contrast is reduced equally for modulated- and static-F0 vowels (Fig. 4C and F).

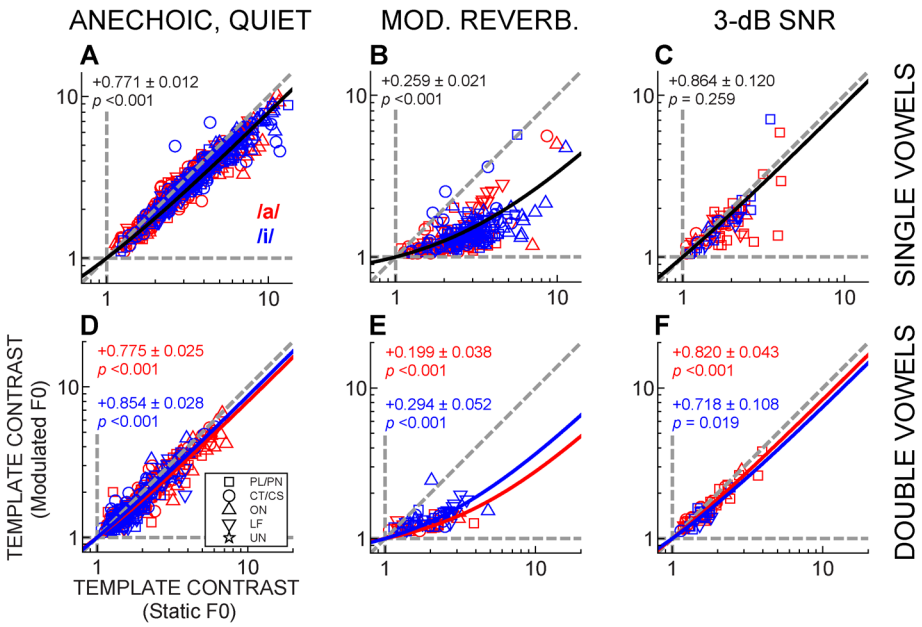


Figure 4. Template contrast: modulated-F0 *vs.* static-F0 vowels. *Top row*, responses to single vowels; *Bottom row*, responses to double vowels. Red symbols indicate responses to /a/, and blue symbols /i/. Horizontal and vertical gray dashed lines indicate template contrast of 1. Diagonal gray dashed line indicates equality. Solid lines indicate least-squares linear fits to the significant data, with the fit constrained to (1,1). Text in each panel indicates the linear-regression slope (β_1) ± S.E., and the p-value for a two-tailed t-test with the null hypothesis H_0: $\beta_1 = 1$. For clarity, only those data from responses with significant contrast (p<0.01) for both static *and* modulated vowels are displayed.

Based on template contrast calculated from double-vowel responses, we computed a "periodicity-tagged" discharge rate for each vowel of the concurrent-speech mixture (Fig. 5). In the across-BF profiles of periodicity-tagged rate, there are peaks and troughs corresponding to the formant structure of each vowel under quiet, anechoic listening conditions. This is the case with both static- and modulated-F0 double vowels (Fig. 5A and D). With reverberation, there are no formant related peaks and troughs remaining in the modulated-F0 case (Fig. 5E), although a clear formant-related pattern of periodicity-tagged discharge rate remains in the static-F0 case in reverberation (Fig. 5B). Information about higher formants is degraded in the presence of noise; however, formant-related peaks in periodicity-tagged discharge rate remain in the first-formant region for both static- and modulated-F0 vowels (Fig. 5C and F).

To quantify the "segregation" of the two vowels of the double-vowel mixture by the periodicity-tagged firing-rate profile across BF, we computed the normalized Euclidean distance between the /a/ and /i/ profiles in response to /a,i/ as (Fig. 6),

$$d_{(/a/,/i/)} = \frac{\sum_{j=1}^{n}\left(\bar{R}_{/a/,j} - \bar{R}_{/i/,j}\right)^2}{\sqrt{n}}$$

The double-vowel sounds are well segregated on the basis of the periodicity-tagged discharge rate statistic, except when F0 is modulated and the sounds are heard in "moderate" or "strong" reverberation (Fig. 6A).

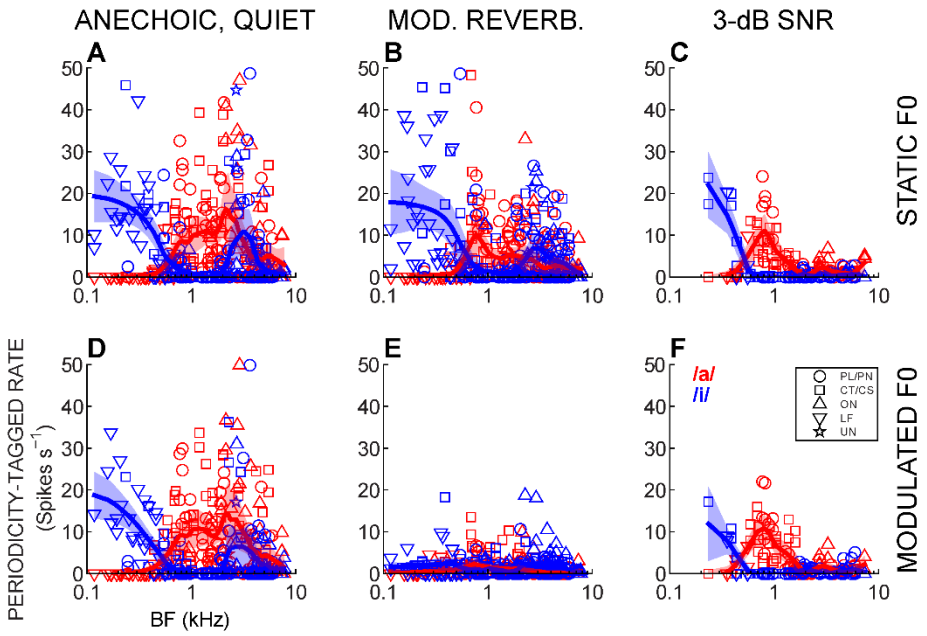


Figure 5. Periodicity-tagged firing rate *vs.* BF. *Top row*, responses to static-F0 vowels; *Bottom row*, responses to modulated-F0 vowels. Solid lines, *lowess* smoothing fits; shaded areas, 95% confidence intervals.

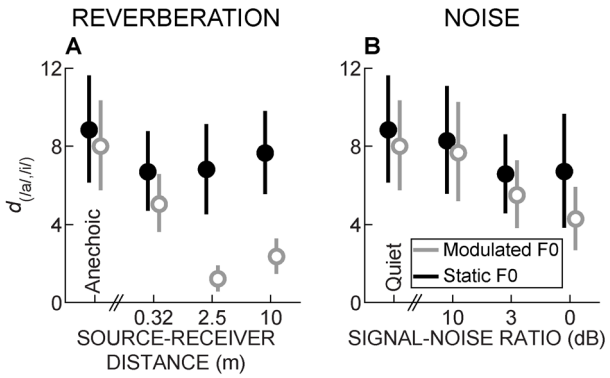


Figure 6. Euclidean distance between /a/ and /i/ periodicity-tagged spike-rate *vs.* BF profiles, calculated from the double-vowel responses.

4 Discussion

The physical effects of noise and reverberation have some similarities: small changes to the magnitude and phase of each component of a complex sound. The impulse response of a room is a linear filter. Therefore, for a steady-state source no new frequency components are added to a reverberant acoustic signal. Contrast this with additive noise. Each indirect sound component in a reverberant space adds to the direct sound at the receiver with essentially random phase, reducing the depth of temporal-envelope modulation at the output of cochlear band-pass filters (Sabine, 1922; Sayles & Winter, 2008; Sayles et al., 2015; Slama & Delgutte, 2015). Perceptually, noise and reverberation can both decrease speech intelligibility (e.g., Nabelek, 1993; Payton et al., 1994), and can be particularly troublesome for cochlear-implant listeners (e.g., Qin & Oxenham, 2005). However, error patterns differ substantially between noisy and reverberant spaces (Nabelek, 1993; Assmann & Summerfield, 2004). Moreover, normal-hearing listeners have remarkably good speech understanding in moderately reverberant spaces (e.g., Poissant et al., 2006).

Our results can be understood in terms of the effects of reverberation on the acoustic temporal envelope. For static-F0 vowels, the reduction in envelope modulation due to reverberant energy is evident in higher-BF units (Fig. 3A and B; Fig. 5A and B). The addition of frequency modulation to the vowels has a dramatic effect on template contrast in reverberation, due to spectral smearing: i.e., monaural decorrelation. Noise impairs neural segregation of concurrent vowels independent of intonation pattern, but in a BF-dependent manner. Similar to other studies of noise-masked single-vowel coding (e.g., Delgutte & Kiang, 1984; May et al., 1998), this is the consequence of higher-BF filters passing more total noise power than low-BF filters. The within-band signal-to-noise ratio is therefore much lower for high-BF units. The negative spectral tilt of the vowels, masked by flat-spectrum noise, imposes a similar frequency-dependency on signal-to-noise ratio after cochlear filtering.

The differential effects of noise and reverberation on speech representations in the early stages of brain-stem neural processing are in clear contrast to noise- and reverberation-invariant representations of speech in auditory cortex (Mesgarani et al., 2014). Perhaps one clue to the neural underpinnings of robust speech understanding in challenging acoustic environments is compensation for the effects of reverberation on coding of temporal-envelope modulation in the inferior colliculus, based on sensitivity to inter-aural correlation (Slama & Delgutte, 2015). What is clear from our data is that neurophysiological effects of room reverberation cannot simply be assimilated to those of broadband noise for signal detection and discrimination.

Acknowledgements

Funded by a grant from the BBSRC (IMW), and a University of Cambridge MD/PhD scholarship (MS).

References

Assmann, P. F., & Summerfield, A. Q. (2004). The perception of speech under adverse conditions. In S. Greenberg, W. A. Ainsworth, R. R. Fay & A. Popper (Eds.), *Speech Processing in the Auditory System*. New York: Springer.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the Perceptual Separation of Simultaneous Voices. *Journal of Phonetics, 10*(1), 23-36.

Culling, J. F., Summerfield, Q., & Marshall, D. H. (1994). Effects of Simulated Reverberation on the Use of Binaural Cues and Fundamental-Frequency Differences for Separating Concurrent Vowels. *Speech Communication, 14*(1), 71-95.

Culling, J. F., Hodder, K. I., & Toh, C. Y. (2003). Effects of reverberation on perceptual segregation of competing voices. *Journal of the Acoustical Society of America, 114*(5), 2871-2876.

Delgutte, B., & Kiang, N. Y. S. (1984). Speech Coding in the Auditory-Nerve. V. Vowels in Background-Noise. *Journal of the Acoustical Society of America, 75*(3), 908-918.

Joris, P. X., Louage, D. H., Cardoen, L., & van der Heijden, M. (2006). Correlation index: a new metric to quantify temporal coding. *Hearing Research, 216-217*, 19-30.

Keilson, S. E., Richards, V. M., Wyman, B. E., & Young, E. D. (1997). The representation of concurrent vowels in the cat anaesthetized ventral cochlear nucleus: Evidence for a periodicity-tagged spectral representation. *Journal of the Acoustical Society of America, 102*(2), 1056-1070.

Larsen, E., Cedolin, L., & Delgutte, B. (2008). Pitch representations in the auditory nerve: two concurrent complex tones. *Journal of Neurophysiology, 100*(3), 1301-1319.

May, B. J., Prell, G. S., & Sachs, M. B. (1998). Vowel representations in the ventral cochlear nucleus of the cat: effects of level, background noise, and behavioral state. *Journal of Neurophysiology, 79*(4), 1755-1767.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America, 111*(18), 6792-6797.

Nabelek, A. K. (1993). Communication in noisy and reverberant environments. In G. A. Stubebaker & I. Hochberg (Eds.), *Acoustical Factors Affecting Hearing Aid Performance*. Needham Heights, USA: Allyn and Bacon.

Palmer, A. R. (1990). The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers. *Journal of the Acoustical Society of America, 88*(3), 1412-1426.

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America, 95*(3), 1581-1592.

Poissant, S. F., Whitmal, N. A., 3rd, & Freyman, R. L. (2006). Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *Journal of the Acoustical Society of America, 119*(3), 1606-1615.

Qin, M. K., & Oxenham, A. J. (2005). Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification. *Ear and Hearing, 26*, 451-460.

Sabine, W. C. (1922). *Collected papers on acoustics*. Cambridge, MA: Harvard University Press.

Sayles, M., & Winter, I. M. (2008). Reverberation challenges the temporal representation of the pitch of complex sounds. *Neuron, 58*, 789-801.

Sayles, M., Stasiak, A., & Winter, I. M. (2015). Reverberation impairs brainstem temporal representations of voiced vowel sounds: challenging "periodicity-tagged" segregation of competing speech in rooms. *Frontiers in Systems Neuroscience, 8*, 248.

Slama, M. C., & Delgutte, B. (2015). Neural coding of sound envelope in reverberant environments. *The Journal of Neuroscience, 35*(10), 4452-4468.

Chapter 42

Audio Visual Integration with Competing Sources in the Framework of Audio Visual Speech Scene Analysis

Ganesh Attigodu Chandrashekara, Frédéric Berthommier, Jean-Luc Schwartz
GIPSA-Lab, Speech and Cognition Department, Grenoble University, France

Contact authors: GAC, ganesh.attigodu@gipsa-lab.grenoble-inp.fr; JLS, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

Abstract

We introduce "Audio-Visual Speech Scene Analysis" (AVSSA) as an extension of the two-stage Auditory Scene Analysis model towards audiovisual scenes made of mixtures of speakers. AVSSA assumes that a coherence index between the auditory and the visual input is computed prior to audiovisual fusion, enabling to determine whether the sensory inputs should be bound together. Previous experiments on the modulation of the McGurk effect by audiovisual coherent vs. incoherent contexts presented before the McGurk target have provided experimental evidence supporting AVSSA. Indeed, incoherent contexts appear to decrease the McGurk effect, suggesting that they produce lower audiovisual coherence hence less audiovisual fusion. The present experiments extend the AVSSA paradigm by creating contexts made of competing audiovisual sources and measuring their effect on McGurk targets. The competing audiovisual sources have respectively a high and a low audiovisual coherence (that is, large vs. small audiovisual comodulations in time). The first experiment involves contexts made of two auditory sources and one video source associated to either the first or the second audio source. It appears that the McGurk effect is smaller after the context made of the visual source associated to the auditory source with less audiovisual coherence. In the second experiment with the same stimuli, the participants are asked to attend to either one or the other source. The data show that the modulation of fusion depends on the attentional focus. Altogether, these two experiments shed light on audiovisual binding, the AVSSA process and the role of attention.

Keywords: Audio Visual Binding, Audio-Visual Speech Scene Analysis, Auditory Speech Analysis, McGurk effect, Attention.

1 Introduction

This paper is focused on a tentative fusion between two separate concepts: Auditory Scene Analysis and Audio-Visual fusion in speech perception.

Auditory Scene Analysis introduced the principle of a two-stage process in the auditory processing of complex auditory scenes with competing sources (Bregman, 1990). A first stage would involve segmenting the scene into auditory elements, which would be segregated or grouped in respect to their common source, either by bottom-up innate primitives or by learnt

top-down schemas. Decision and formation of a final percept would be done at a second later stage.

Audio-Visual fusion in speech perception refers to the well-known fact that speech perception involves and integrates auditory and visual cues, as shown in various paradigms such as speech in noise (Sumby & Pollack, 1954; Erber, 1969) or the perception of conflicting stimuli (the so-called McGurk effect, McGurk & MacDonald, 1976; also see Tiippana, 2014)

Since a pioneer proposal by Berthommier (2004), our group proposed that auditory scene analysis and multisensory interactions in speech perception should be combined into a single "Audio-Visual Speech Scene Analysis" (AVSSA) process. The basic claim is that the two-stage analysis-and-decision process as work in ASA should be extended to audiovisual speech scenes made of mixtures of auditory and visual speech sources. A first audiovisual binding stage would involve segmenting the scene into audiovisual elements, which should be segregated or grouped in respect to their common multisensory speech source, either by bottom-up *audiovisual primitives* or by learnt top-down *audiovisual schemas*. This audiovisual binding stage would control the output of the later decision stage, and hence intervene on the output of the speech-in-noise or McGurk paradigms.

To provide evidence for this "binding and fusion" AVSSA process, Nahorna et al. (2012, 2015) showed that the McGurk effect can be significantly and strongly reduced by an audiovisual context made of a few seconds of incoherent material (sounds and images coming from different speech sources) presented before the McGurk target (audio "ba" plus video "ga"): the target, classically perceived as "da" in the McGurk effect, was more often perceived as "ba", suggesting a decreased weight of the visual input in the fusion process. The interpretation was that the incoherent context resulted in an "unbinding" effect decreasing the visual weight and hence diminishing the McGurk effect. This modulation of the McGurk effect through incoherent contexts was further extended to speech in noise (Ganesh et al., 2013), and a possible neurophysiological correlate of the binding/unbinding process was provided in an EEG experiment (Ganesh et al., 2014).

However, these studies were based on audiovisual scenes that never implied competing sources. The objective of the present study was to test the "binding and fusion" AVSSA process in scenes including competition between audiovisual sources. For this aim, we generated two audiovisual sources, one made of a sequence of isolated syllables, and the other one made of a sequence of sentences. We prepared two kinds of combinations, with the same auditory content (mixing the two audio sources, syllables and sentences) and two different video contents, either the syllables or the sentences. These two combinations ("Video syllables" & "Video sentences") were used as the context in a McGurk experiment. We hypothesized that since syllables correspond to stronger audiovisual modulations in time and hence stronger audiovisual coherence than sentences, the association between the visual input and the corresponding auditory input would be stronger for syllables than for sentences. Hence the coherence of the audiovisual context would be stronger for syllables, and it would lead to a larger visual weight, hence more McGurk effect than with visual sentences. Furthermore, the introduction of a competition between sources made it possible to introduce attention factors in the paradigm, and we tested whether the attentional focus put by the participants on either syllables or sentences would play a role in the AVSSA process.

2 Method

2.1 Participants

The study involved twenty-nine French participants without hearing or vision problems (22 women and 7 men; 27 right-handed and 2 left handed; mean age= 29.2 years; SD=10. 4 years), who all gave informed consent to participate in the experiment.

2.2 Stimuli

The stimuli were similar to those of the previous experiment by (Nahorna et al., 2015) with suitable modification in the paradigm. They were prepared from two sets of audiovisual material, a "syllables" material and a "sentences" material, produced by a French male speaker, with lips painted in blue to allow precise video analysis of lip movements (Lallouache, 1990). The whole experiment consisted of two types of contexts followed by a target.

Target

The target was either a congruent audiovisual "ba" syllable ("ba-target" in the following), serving as a control – or an incongruent McGurk stimulus with an audio "ba" mounted on a video "ga" ("McGurk target" in the following).

Context

There were two types of contexts i.e. "Video syllables" and "Video sentences". In both contexts, the set of audio stimuli was the same. It consisted of a sequence of 2 or 4 syllables (A-syl-2 or A-syl-4) randomly extracted from the set "pa," "ta," "va," "fa," "za," "sa," "ka," "ra," "la," "ja," "cha," "ma," or "na," mixed with a portion of a random set of sentences with the adequate duration (A-sent-2 or A-sent-4). The 2- vs. 4-syllable duration was selected from earlier experiments by Nahorna et al. (2015), showing that the effect of incoherent context was maximal (maximal reduction of the McGurk effect) for short 2-syllable contexts and slightly less for longer 4-syllable contexts. The visual components of the context were the visual stream associated with either the auditory syllables (V-syl-2 or V-syl-4) or the auditory sentences (V-sent-2 or V-sent-4). Therefore, in the "Video syllables" contexts, there was an audiovisual "syllables" source competing with an audio "sentences" source, while in the "Video sentences" contexts, there was an audiovisual "sentences" source competing with an audio "syllables" source (Fig. 1). A 200 ms fading transition stimulus (five images) was implemented between context and target to ensure continuity between images.

There were altogether 120 stimuli with four times more "McGurk" than "Ba" targets (serving as controls), and with the same number of occurrences of the V-syl-2, V-syl-4, V-sent-2 and V-sent-4 contexts (6 occurrences each for "Ba" targets, 24 occurrences each for McGurk targets). The 120 stimuli were presented in a random order and concatenated into a single 7-minutes film.

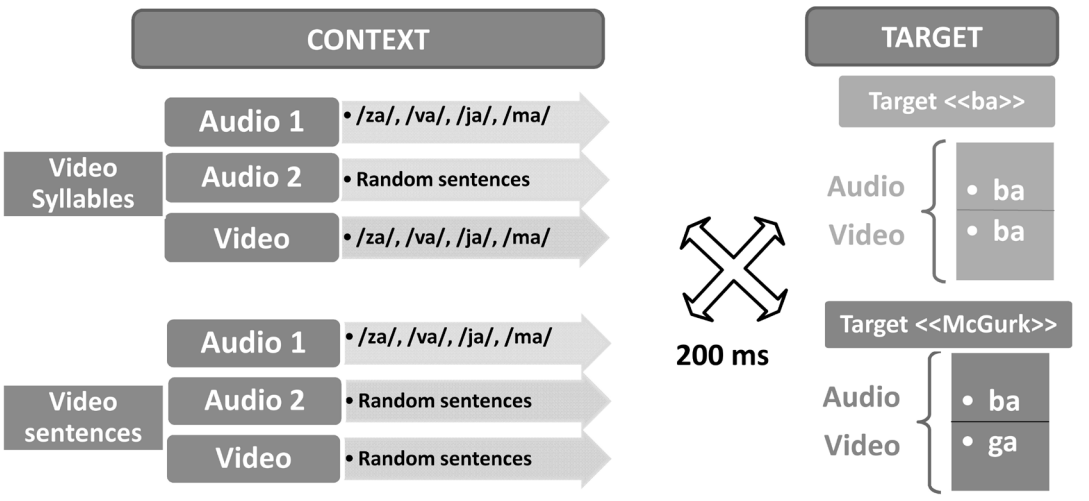


Figure 1. Description of the audiovisual material

2.3 Procedure

The study included two consecutive experiments, Exp. A. followed by Exp. B (always in this order). In Exp. A, the participants were involved in a monitoring paradigm in which they were asked to constantly look at the screen and monitor for possible "ba" or "da" targets by pressing an appropriate key, as in Nahorna et al. (2012, 2015). In Exp. B the monitoring "ba" vs. "da" task remained the same (with a different order of the 120 stimuli in the film), but in addition, specific instructions were given to participants, either to put more attention to syllables ("Attention syllables") or to put more attention to sentences ("Attention sentences"). The order of the "Attention syllables" and "Attention sentences" conditions was counterbalanced between the participants. To increase the efficiency of the attentional demand, participants were informed them that they would be questioned on the content of either the "syllables" or the "sentences" material at the end of the experiment. A practice session was provided for all of them and most of the participants were indeed able to recall specific syllables or words. The films were presented on a computer monitor with high-fidelity headphones set at a comfortable fixed level.

2.4 Processing of responses

Response time was computed in reference to the acoustic onset of the burst of the "b" in the target syllable, discarding values higher than 1200 ms or lower than 200 ms. "ba" and "da" responses were taken into account only when they occurred within this time window [200-1200ms] and two different responses inside the time window were also discarded. Finally, for each participant and each condition of context and target (and attention in Exp. B), a global score of "ba" responses was calculated as the percentage of "ba" responses divided by the sum of "ba" and "da" responses to the target, and a mean response time was calculated as the average of response times for all the responses to the target.

3 Results

First of all, the mean percentage of "ba" scores for McGurk targets over all conditions in Exp. A was computed for each subject, and participants providing mean scores larger than 95% or less than 5 % were discarded, considering that these subjects provided either too strong or too low McGurk effects to enable binding modulations to be displayed. This resulted in discarding 8 out of 29 participants. All further analyses for both Exp. A and B will hence concern only the 21 remaining subjects. As expected, the global score (percentage of "ba" responses relative to "ba" + "da" responses) for all control "ba" targets was close to 100 % in all conditions in both experiments. Therefore, from now on we will concentrate on McGurk targets.

3.1 On the role of context type without explicit attention focus (Exp. A)

Percentages of "ba" responses to McGurk targets in Exp. A (without explicit attentional focus) are displayed on the left part of Fig. 2. A two-factor repeated measures ANOVA with *context type* ("Video syllables" vs. "Video sentences") and *context duration* (2- vs. 4-syllables) as the independent variables was administered on these percentages (applying Greenhouse-Geisser correction when applicable). The effect of *context type* is significant [F (1, 20) =34. 65, p<0.001], with a higher McGurk effect (10 % less "ba" responses) with the "Video syllables" context. This is in line with our prediction that audiovisual coherence is higher in the "Video syllables" condition, leading to a higher binding level, a larger visual weight and hence a larger number of McGurk fusion ("da" responses). *Context duration* displayed no significant effect on "ba" scores, either in isolation or in interaction with *context type*.

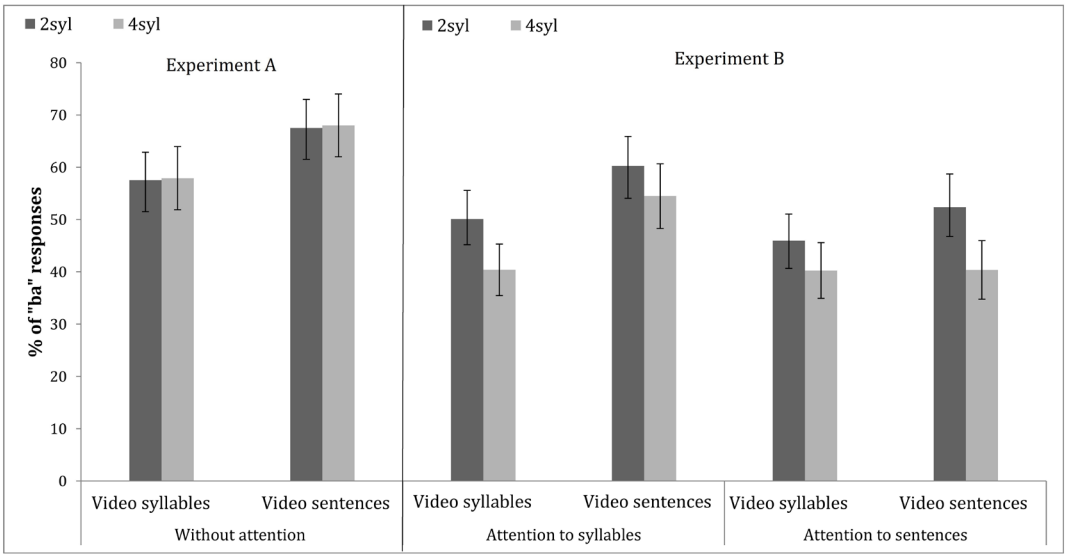


Figure 2. The percentage of "ba" responses (relative to the total number of "ba" or "da" responses) for "McGurk" targets, in the "Video syllables" vs. "Video sentences" contexts in Experiment A and Experiment B.

3.2 On the interaction between context type and attention focus (Exp. B)

Percentages of "ba" responses to McGurk targets in Exp. B (involving explicit attention towards one or the other source) are displayed on the right part of Fig. 2. A repeated-measures ANOVA was administered on these percentages with three factors, *context type* ("Video syllables" vs. "Video sentences"), *context duration* (2- vs 4-syllables) and *attention* ("Attention syllables" vs. "Attention sentences") by applying Greenhouse-Geisser correction when applicable.

The effect of *context type* [F (1, 20) =11. 91, p<0.001] is significant, as in Exp. A: video syllables produce more McGurk than video sentences. Contrary to Exp. A, the effect of *context duration* [F (1, 20) =33. 86, p<0.001] is also significant, with no interaction with *context type*. The *attention* factor alone is not significant, but its interaction with context type is significant [F (1, 20) =11.07, p<0.005]. Post-hoc analyses with Bonferroni corrections show that while there is no significant difference between the two *attention* conditions for the "Video syllables" context type, there is a difference for the "Video sentences" condition, with a lower "ba" percentage (a higher McGurk effect) in the "Attention sentence" condition. Interestingly, while the "ba" percentage is higher for the "Video sentences" than for the "Video syllables" condition when attention is put in syllables; there is no more significant difference when attention is put on sentences.

Finally, the three-way interaction between context type, context duration and attention is significant [F (1, 20) =6. 51, p<0.05], with a larger difference between durations from the "Video syllables" to the "Video sentences" condition in the "Attention sentences" than in the "Attention syllables" condition.

3.3 Response time

The results are consistent with the previous findings (Nahorna et al., 2012) in which response times were larger for McGurk targets, independently on context. In both experiments and in all contexts, the processing of "ba" responses was indeed quicker compared to McGurk responses. Two-way repeated-measures ANOVA on *target* and *context type* in Exp. A displays an effect of target [70ms quicker response for "ba" targets, F (1, 20) =14.25, p<0.005] and no effect of context or any interaction effect. A three-way repeated-measures ANOVA on *condition type*, *attention* and *target* in Exp. B displays once again an effect of target [80ms quicker response for

"ba" targets, F (1, 20) =4.47, p<0.05] and no other significant effect of other factors, alone or in interaction.

4 Discussion

Audiovisual fusion has long been considered as an automatic fusion process (e.g. Massaro, 1987). However Exp. A. confirms that the contextual stimulus may modulate fusion as in our previous experiments (Nahorna et al., 2012; Ganesh et al., 2013; Nahorna et al., 2015) and extends the concept to the case of competing sources.

This supports the AVSSA hypothesis, in which a first speech scene analysis process would group together the adequate audiovisual pieces of information and estimate the degree of audiovisual coherence. The effect of context type (larger McGurk effect in the "Video syllables" condition) could be due to the differences in audiovisual correlations for syllables and sentences. Indeed, correlation analysis between audio (full band envelope) and video (mouth opening area) material for syllables and sentences provides a mean correlation value of 0.59 for "Video syllables" and 0.10 for "Video syllables" (Fig. 3). Another factor could increase binding with syllables, i.e. the presence of a streaming mechanism in which the syllabic target would be associated to the syllables stream rather than to the sentences stream.

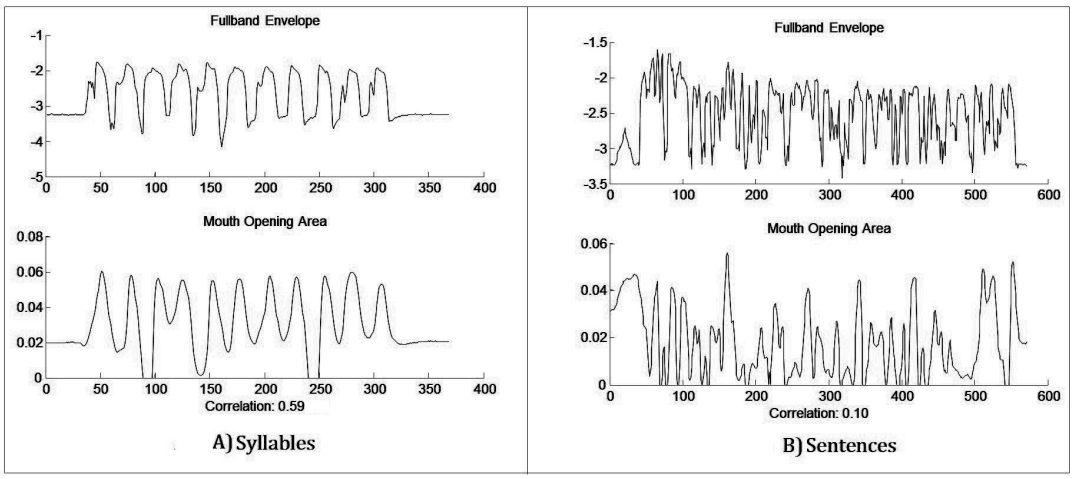


Figure 3. The correlation values between audio and video for syllables and sentences.

A number of recent experiments pointed the role of general attentional mechanisms able to globally decrease the amount of fusion (Tiippana et al., 2004; Alsius et al., 2007; Navarra et al., 2010; Alsius et al., 2014). Experiment B shows that attentional mechanisms may intervene at the level of single audiovisual sources in an audiovisual speech scene, selectively increasing or decreasing the amount of fusion depending on the coherence of the attended source. Interestingly, attention intervened only for "Video sentences". Our interpretation is that binding could be pre-attentive for syllables, because of their salient audiovisual comodulations making them pop out as strong bottom-up audiovisual primitives. On the contrary, the coherence of audiovisual sentences is low hence the attentional focus could enhance audiovisual top-down schemas increasing binding.

These two studies provide confirmation and development to the view that audiovisual fusion in speech perception includes a first stage of audiovisual speech scene analysis. A number of previous studies suggested that the presentation of a visual stream can enhance segregation by affecting primary auditory streaming (Rahne et al., 2007; Marozeau et al., 2010; Devergie et al., 2011) or that visual cues can improve speech detection and cue extraction (Grant & Seitz, 2000; Kim & Davis, 2004; Schwartz et al., 2004; Alsius & Munhall, 2013); though see some

contradictory studies highlighting cases where unimodal perceptual grouping precedes multisensory integration (Sanabria et al., 2005).

Altogether, the "binding stage" in the AVSSA process, in which the coherence between auditory and visual features would be evaluated in a complex scene, provides a central mechanism in order to properly associate the adequate components inside a coherent audiovisual speech source. The present study confirms the importance of this mechanism in which the "binding stage" enables the listener to integrate "similar sources" or segregate "dissimilar sources" in Audio Visual fusion.

Acknowledgments

This project has been supported by Academic Research Community "Quality of life and ageing" (ARC 2) of the Rhône-Alpes Region, which provided a doctoral funding for Ganesh Attigodu Chandrashekara.

References

Alsius, A., & Munhall, K. G. (2013). Detection of audiovisual speech correspondences without visual awareness. *Psychological Science, 24*(4), 423-431. doi: 10.1177/0956797612457378

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research, 183*(3), 399-404. doi: 10.1007/s00221-007-1110-1

Alsius, A., Mottonen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Frontiers in Psychology, 5*, 727. doi: 10.3389/fpsyg.2014.00727

Berthommier, F. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication, 44*(1–4), 31-41. doi: 10.1016/j.specom.2004.10.003

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge: MIT Press, MA.

Devergie, A., Grimault, N., Gaudrain, E., Healy, E. W., & Berthommier, F. (2011). The effect of lip-reading on primary stream segregation. *The Journal of the Acoustical Society of America, 130*(1), 283-291. doi: 10.1121/1.3592223

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12*(2), 423-425.

Ganesh, A. C., Berthommier, F., & Schwartz, J.-L. (2013). *Effect of context, rebinding and noise on audiovisual speech fusion.* Paper presented at the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France.

Ganesh, A. C., Berthommier, F., Vilain, C., Sato, M., & Schwartz, J.-L. (2014). A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Frontiers in Psychology, 5*, 1340. doi: 10.3389/fpsyg.2014.01340

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197-1208. doi: 10.1121/1.1288668

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. *Speech Communication, 44*(1–4), 19-30. doi: 10.1016/j.specom.2004.09.008

Lallouache, M. T. (1990). *Un poste 'visage-parole.' Acquisition et traitement de contours labiaux (A 'face-speech' workstation. Acquisition and processing of labial contours).* Paper presented at the Proceedings of the eighteenth Journées d'Etudes sur la Parole, Montréal, QC.

Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N., & Blamey, P. J. (2010). The Effect of Visual Cues on Auditory Stream Segregation in Musicians and Non-Musicians. *PLoS ONE, 5*(6), e11297. doi: 10.1371/journal.pone.0011297

Massaro, D. W. (1987). *Speech Perception by Ear and Eye*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. [10.1038/264746a0]. *Nature, 264*(5588), 746-748.

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America, 132*(2), 1061-1077. doi: 10.1121/1.4728187

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2015). Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America, 137*(1), 362-377. doi: 10.1121/1.4904536

Navarra, J., Agn, #232, Alsius, s., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Inf. Fusion, 11*(1), 4-11. doi: 10.1016/j.inffus.2009.04.001

Rahne, T., Bockmann, M., von Specht, H., & Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research, 1144*, 127-135. doi: 10.1016/j.brainres.2007.01.074

Sanabria, D., Soto-Faraco, S., Chan, J., & Spence, C. (2005). Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. *Neuroscience Letters, 377*(1), 59-64. doi: 10.1016/j.neulet.2004.11.069

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition, 93*(2), B69-78. doi: 10.1016/j.cognition.2004.01.006

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*(2), 212-215. doi: 10.1121/1.1907309

Tiippana, K. (2014). What is the McGurk effect? [Opinion]. *Frontiers in Psychology, 5*. doi: 10.3389/fpsyg.2014.00725

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457-472. doi: 10.1080/09541440340000268

Chapter 43

Relative Pitch Perception and the Detection of Deviant Tone Patterns

Susan L. Denham[1], Martin Coath[1], Gábor P. Háden[2], Fiona Murray[1] and István Winkler[2,3]

[1] *Cognition Institute and School of Psychology, Plymouth University, Plymouth, UK*
[2] *Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary*
[3] *Institute of Psychology, University of Szeged, Szeged, Hungary*

Contact author: SLD, sdenham@plymouth.ac.uk

Abstract

Most people are able to recognise familiar tunes even when played in a different key. It is assumed that this depends on a general capacity for relative pitch perception; the ability to recognise the pattern of inter-note intervals that characterises the tune. However, when healthy adults are required to detect rare deviant melodic patterns in a sequence of randomly transposed standard patterns they perform close to chance. Musically experienced participants perform better than naïve participants, but even they find the task difficult, despite the fact that musical education includes training in interval recognition.

To understand the source of this difficulty we designed an experiment to explore the relative influence of the size of within-pattern intervals and between-pattern transpositions on detecting deviant melodic patterns. We found that task difficulty increases when patterns contain large intervals (5-7 semitones) rather than small intervals (1-3 semitones). While task difficulty increases substantially when transpositions are introduced, the effect of transposition size (large vs small) is weaker. Increasing the range of permissible intervals to be used also makes the task more difficult. Furthermore, providing an initial exact repetition followed by subsequent transpositions does not improve performance. Although musical training correlates with task performance, we find no evidence that violations to musical intervals important in Western music (i.e. the perfect fifth or fourth) are more easily detected. In summary, relative pitch perception does not appear to be conducive to simple explanations based exclusively on invariant physical ratios.

Keywords: relative pitch perception, oddball paradigm, translation-invariant perception

1 Introduction

Most people easily recognise well known melodies even when they are transposed to a different key. The invariant property of transposed melodies is the preserved pitch ratio relationship between notes of the melody; i.e. pitch intervals of the melody remain the same despite changes in absolute pitch. For this reason, it is assumed that the ability to recognise pitch relationships (relative pitch perception) is rather robust and commonly found in the population. Recognition of preserved pitch interval patterns irrespective of absolute pitch is an auditory example of translation-invariant object perception (Griffiths & Warren, 2004; Kubovy & Van Valkenburg, 2001; Winkler, Denham, & Nelken, 2009).

The robustness of the ability to recognise tone patterns has been supported by recent findings showing that listeners can detect random tone patterns very quickly (after ca. 1.5 repetitions) within rapidly presented tone sequences, even if the patterns are quite long (up to 20 tones in a pattern) (Barascud, 2014). The human brain is also sensitive to pattern violations, with regular to random transitions (Chait, Poeppel, de Cheveigne, & Simon, 2007) being detected within about 150 ms (~3 tones) from deviation onset (Barascud, 2014). However, in these examples tone patterns were always repeated exactly, i.e. without transposition, so it is not clear whether listeners were remembering absolute pitch sequences or relative pitch relationships.

In support of the assumed generality of relative pitch perception, it has been shown that violations of transposed pitch patterns elicit discriminative brain responses in neonates (Stefanics et al., 2009) and young infants (Tew, Fujioka, He, & Trainor, 2009). So it is surprising that relative pitch perception can be rather poor (e.g. see Foster & Zatorre, 2010; McDermott, Keebler, Micheyl, & Oxenham, 2010), especially if contour violations and tonal melodies are excluded (Dowling, 1986). McDermott et al. (2010), commenting on the poor pitch interval discrimination threshold they found, suggested that the importance of pitch as an expressive musical feature may rest more on an ability to detect pitch differences between tones, rather than an ability to recognise complex patterns of pitch intervals.

Some years ago, in a pilot experiment we noticed that an oddball interval (e.g. a tone pair separated by 7 semitones) did not pop out as expected within a randomly transposed series of standard intervals (e.g. 3 semitones). We subsequently ran a series of experiments in which we maintained a standard pitch contour, but varied the number of repetitions of the standard phrase (2 or 3), the number of tones in a phrase (2-6), the size of the deviance (1-3 semitones), and the tonality of the short melodies (Coath, 2008). Most listeners, including those with musical education, found it very difficult to detect an oddball melodic phrase in a sequence of randomly transposed standard phrases, performing close to chance. The source of the surprising difficulty of the task was not clarified by this experiment, as the variables tested only weakly influenced performance. Here we report another attempt to discover what makes this task so hard.

Consistent with Gestalt grouping principles (Köhler, 1947), auditory streaming experiments show that featural separation (such as pitch differences) promote segregation and conversely that featural similarity promotes integration (Bregman, 1990; Moore & Gockel, 2012). It is also known that within-stream (within-pattern) comparisons are far easier to make than between stream comparisons; (e.g. Bregman, 1990; Micheyl & Oxenham, 2010). Therefore, we hypothesized that if the standard pattern satisfied Gestalt grouping principles and could thus be more easily grouped, this would facilitate pattern comparisons, and that deviations within such patterns would be easier to detect. Another possibility is that confusion between within-pattern intervals and between-pattern transpositions may make individual patterns less distinctive, and so increase the task difficulty. Therefore, we also investigated the effects of transposition size and interactions between transposition size and within-phase intervals. Finally, the predictive coding account of perception (Friston, 2005) suggests that the precision with which perceptual discriminations can be made is inversely related to stimulus variance, suggesting that task difficulty would increase with variance of standard phrase pitch intervals.

Our specific hypotheses were:

1. Small within-pattern intervals will promote grouping and thus improve performance (Gestalt proximity/similarity);
2. Small transpositions, especially when within-pattern intervals are large, may make individual patterns less distinctive, and thus impair performance;
3. Exact repetitions with no transposition will result in very good performance;
4. One exact repeat (i.e. pattern 1 == pattern 2) before introducing transpositions may allow a better pattern representation to be built and used as a template for subsequent patterns, and so improve task performance;

5. Smaller variance in the intervals within a pattern (either only small or only large intervals) will increase the predictability of the pattern and allow the formation of a more precise representation of the pattern. Therefore, task performance will decrease with increasing interval variance.
6. Musical training and experience will facilitate task performance.

2 Methods

The study was approved by the ethical review board of Plymouth University. Participants either received credits in a university course for their participation, or volunteered to take part.

2.1 Participants

Data were collected from 54 participants in total (32 females, 22 males; age range 19-65 years, median 20.5 years). The majority were undergraduate Psychology students at Plymouth University. Additional participants recruited from a doctoral programme and the University orchestra. All participants confirmed they had normal hearing. Details of musical training (years of formal tuition) and playing experience (years playing) were recorded for each participant. Four participants' data were excluded from the analysis as they achieved less than 30% in at least one experimental block (chance level being 50%), suggesting that they may not have understood the task correctly.

2.2 Materials

The experiment was conducted using a bespoke Matlab programme. Participants listened to the stimuli using Sennheiser HD215 headphones, individually adjusted to a comfortable sound level during the initial practice trial. The absolute level selected by each participant was not recorded.

Stimuli

Each trial consisted of four patterns separated by 700ms silence, and each pattern consisted of six tones. Three of the patterns had the same sequence of pitch intervals (standard pattern); the last pitch interval of either the final or the penultimate pattern of the trial deviated from the other three. A different standard pattern was delivered on each trial and no pattern was used more than once in the experiment. Patterns were generated by randomly selecting a set of five intervals, with the restrictions that each interval should only occur once within a pattern, and two intervals with same magnitude but opposite sign should not follow each other immediately in the sequence (to prevent the occurrence of repeated tones in the pattern).

All tones making up the pitch sequences were harmonic complexes, consisting of the first four harmonics of the nominal pitch, exponentially decreasing in amplitude (1:1/2:1/4:1/8) to give an oboe-like timbre. Tone duration was 110ms, with 5ms onset and offset linear ramps and 40ms silence between tones, giving a tone onset to onset interval of 150ms. Deviant intervals were always four semitones. Since standard pattern intervals were chosen from the set {1,2,3,5,6,7 semitones}, depending on the condition (see Table 1), the difference between the standard and the deviant pitch interval was always 1, 2 or 3 semitones. The first tone of the first pattern always had a pitch of 450Hz. To avoid the use of pitches which may not be clearly audible to everyone despite reporting normal hearing, all pitches were restricted to lie between 100 and 3200 Hz.

The experiment consisted of one practice block and seven test blocks, each distinguished by the set of intervals used, as detailed in Table 1. Intervals were nominally divided into two sets: *small* {1,2,3} semitones, and *big* {5,6,7} semitones.

The practice block consisted of 10 trials. The first four were very easy with no transpositions and small within pattern intervals. The next four were slightly harder with two exact repeats of

the pattern before two transpositions, with small within-pattern intervals and small transpositions. The final two examples were similar to trials in block 3 with small within-pattern intervals and small transpositions. Participants were given feedback after each trial (the response button briefly turned green for correct and red for incorrect) and a final practice score.

Table 1. Details of the within-pattern and transposition intervals used and the number of trials in each test block.

Block	Within-pattern intervals	Transposition intervals	Number of trials
1	big	none	10
2	big	one exact repeat, then two big transpositions	10
3	small	small	20
4	small	big	20
5	big	small	20
6	big	big	20
7	all: 1,2,3,5,6,7 ST	big	20

Procedure

Participants were required to indicate using two on-screen response buttons (labelled '2nd Last' and 'Last') whether the penultimate or last pattern was different from the rest. They were told that any difference was in the last interval of the pattern.

Participants began by entering their personal details and then continued with the practice block. They were encouraged to repeat the practice block as many times as they needed to familiarize themselves with the task; 1-3 repetitions were judged to suffice in all cases.

Following the practice block, participants were presented with seven test blocks, with no feedback. Blocks as detailed in Table 1 were presented in random order. Once they had completed all the test blocks, participants were presented with a bar graph showing their score in each block. Each 20-trial block took 3-4 minutes to complete and the experiment lasted roughly 30 minutes.

Analysis

In all cases confidence was assessed at the .05 level. Score distributions in each test block were compared against chance using the T-test. The effect of block was assessed using a 1-way ANOVA with all test blocks. The effect of transposition was assessed by contrasting block 1 with the average of blocks 5 and 6. The effect of one exact repetition was assessed by contrasting block 2 with block 6. The effect of variance in interval range was assessed by contrasting block 7 with the average of blocks 4 and 6. The effects of within-phrase intervals and between-phrase transpositions on performance were assessed using a two way ANOVA on data from blocks 3-6. The effect of interval variance was also tested using correlation analysis on data from blocks 3-7. The effect of final interval size of performance was tested using correlation analysis on data from all test blocks. The effect of musical experience were tested using correlation analysis on data from all test blocks. Correlation analysis was carried out using Spearman's correlation coefficient as the data were not normally distributed.

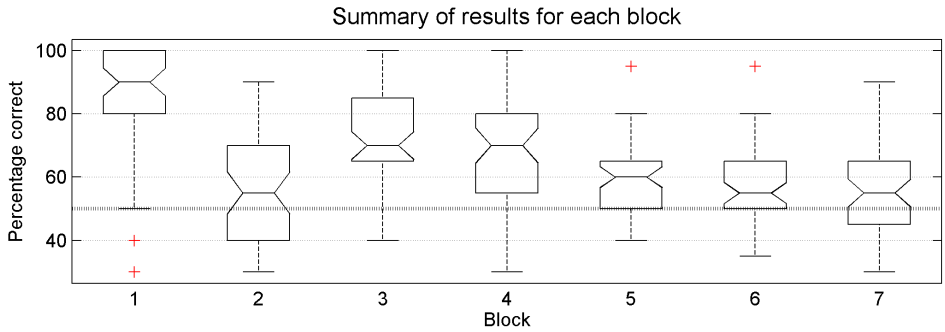


Figure 1. Distribution of percent correct scores in each block across participants.

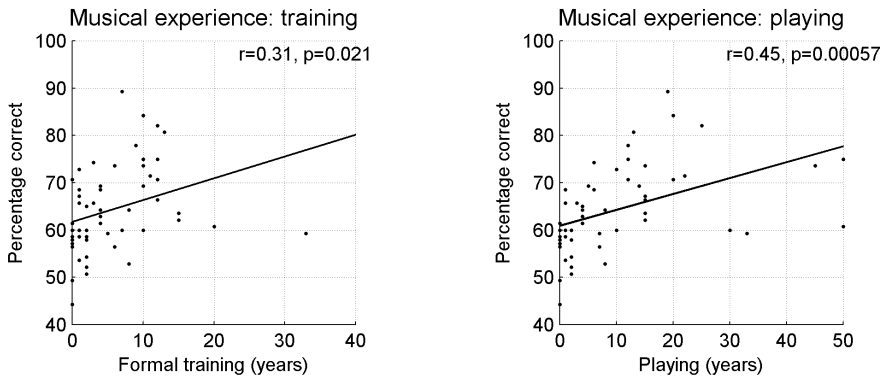


Figure 2. The influence of musical experience on task performance.

3 Results

Figure 1 shows the score distributions for each block for the participants.

Performance in all blocks was found to be significantly different from chance (shown by dotted line in the figure; $p < 0.05$).

There was a main effect of block ($F(6,294) = 41.61$, $p < .001$, $\varepsilon = .790$, partial $\eta^2 = .459$). The effect of transposition (contrasting block 1 with the average of blocks 5 and 6) was significant ($t = -10.36$, $p < .001$). The effect of one exact repetition (contrasting block 2 with block 6) was not significant ($t = .59$, $p = .559$). The effect of variance in interval range (contrasting block 7 with the average of blocks 4 and 6) was significant ($t = 3.20$, $p = .002$). The more detailed trial-level correlation analysis showed performance correlated negatively with the variance of the pattern intervals (correlation coefficient = -.336, $p < .001$)). There was no significant correlation between the magnitude of the final interval and performance (correlation coefficient -0.298, $p = .147$). Posthoc multiple comparison analysis showed performance for musically important final intervals (perfect fourth and fifth, 5 and 7 semitones, respectively) was significantly lower than that for 1 semitone.

The two-way ANOVA assessing the effect of within-pattern and transposition intervals showed a significant main effect of within-pattern intervals ($F(1,49) = 37.45$, $p < .001$, partial $\eta^2 = .433$) and transposition size ($F(1,49) = 12.16$, $p = .001$, partial $\eta^2 = .199$) but their interaction was not significant ($F(1,49) = 1.45$, $p = .235$, partial $\eta^2 = .029$). A posthoc multiple comparison analysis showed that there was a tendency for large transpositions to impair performance more than small transpositions.

Performance correlated positively with musical experience; years of formal training (correlation coefficient = .342, p = .015), as well as years of playing (correlation coefficient = .435, p =.002).

The influence of musical training on task performance is illustrated in Figure 2.

4 Discussion

In this study we investigated some of the potential sources of difficulty in detecting a pattern with a deviant pitch interval amongst transposed repetitions of a standard pattern, a task that is assumed to depend on relative pitch perception. Our results are consistent with a number of previous studies, e.g. McDermott & Oxenham (2008), showing that relative pitch perception may be more limited than is commonly assumed. Performance is best when the standard phrase is repeated exactly with no transpositions (block 1), but falls substantially when transpositions are introduced (block 1 versus the average of blocks 3-7). Without transpositions, the task can be performed by direct comparisons between pitches, rather than using the interval relationships between successive pitches. Performance is not helped by one exact repetition of the standard pattern (block 2 versus block 6). This shows that although listeners may become sensitive to a repeating pattern after only 1.5 repetitions (Barascud, 2014), they are unable to use this pattern for comparison with transposed versions of the pattern.

When patterns are transposed, then performance is best for standard patterns consisting of small intervals. This is consistent with the notion that grouping is promoted by featural similarity, and that representations of phrases consisting of small intervals are more easily formed, suggesting that comparisons between patterns may be facilitated by having a more coherent representation of the standard. With transpositions, large within-pattern intervals make the task very difficult. However, contrary to our hypothesis, large transpositions impaired performance more than small transpositions. This suggests that comparisons between pitch interval patterns are facilitated by proximity in pitch space. Increasing the variance in the pattern intervals, as predicted, impairs performance.

The idea that relative pitch perception depends solely on detecting a pattern of invariant pitch intervals is not supported by our results. Although the invariant property of the patterns in each trial is the sequence of pitch intervals defining the standard, listeners often could not use this information in the current experiment. Our results are compatible with the notion that in constructing object representations, the tolerance of the representation is a function of the variance in the pattern, i.e. increasing variance in object components lead to more permissive representations. This makes sense when the general problem of perceptual categorisation is considered; e.g. the variability of the spoken word.

Relative pitch perception has been likened to translation invariant object recognition in vision (Kubovy & Van Valkenburg, 2001). Interestingly the literature on visual perceptual learning has shown that learning can be surprisingly specific to the precise retinal location of the task stimulus (Fahle, 2005). The most influential model of translation invariant object recognition is the so-called trace model (Stringer, Perry, Rolls, & Proske, 2006), which assumes that this ability actually depends on learning the activity caused by the same stimulus being shown at many different locations; invariant recognition then emerges at a higher level by learning that these different activations are caused by the same object. Perhaps this is what happens when we learn a tune. The categorisation of the tune depends on hearing it at many different pitch levels within a context that provides clear links between the various repetitions (e.g. within the same piece of music, or same social context).

Acknowledgments

We would like to thank Dávid Farkas for help with the statistical analysis. IW was supported by the Hungarian Academy of Sciences (Lendület project, LP36/2012). GPH was supported by a post-doctoral research fellowship of the Hungarian Academy of Sciences.

References

Barascud, N. (2014). *Auditory Pattern Detection.* (PhD thesis), University College London (University of London), London.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Chait, M., Poeppel, D., de Cheveigne, A., & Simon, J. Z. (2007). Processing asymmetry of transitions between order and disorder in human auditory cortex. *J Neurosci, 27*(19), 5207-5214. doi: 10.1523/JNEUROSCI.0318-07.2007

Coath, M., Háden, G.P., Winkler, I., Denham, S.L. (2008). *Are short tone sequences memorable?* Paper presented at the British Society of Audiology Annual Conference, Leeds.

Dowling, W.J. (1986). Context Effects on Melody Recognition: Scale-Step versus Interval Representations. *Music Perception, 3*(3), 281-296.

Fahle, M. (2005). Perceptual learning: specificity versus generalization. *Curr Opin Neurobiol, 15*(2), 154-160. doi: 10.1016/j.conb.2005.03.010

Foster, N. E., & Zatorre, R. J. (2010). A role for the intraparietal sulcus in transforming musical pitch information. *Cereb Cortex, 20*(6), 1350-1359. doi: 10.1093/cercor/bhp199

Friston, K. (2005). A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci, 360*(1456), 815-836. doi: 10.1098/rstb.2005.1622

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nat Rev Neurosci, 5*(11), 887-892. doi: 10.1038/nrn1538

Köhler, W. (1947). *Gestalt psychology: An introduction to new concepts in modern psychology*. New York: Liveright Publishing Corporation.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition, 80*(1-2), 97-126.

McDermott, J. H., Keebler, M. V., Micheyl, C., & Oxenham, A. J. (2010). Musical intervals and relative pitch: frequency resolution, not interval resolution, is special. *J Acoust Soc Am, 128*(4), 1943-1951. doi: 10.1121/1.3478785

McDermott, J. H., & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Curr Opin Neurobiol, 18*(4), 452-463. doi: 10.1016/j.conb.2008.09.005

Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *J Assoc Res Otolaryngol, 11*(4), 709-724. doi: 10.1007/s10162-010-0227-2

Moore, B. C., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philos Trans R Soc Lond B Biol Sci, 367*(1591), 919-931. doi: 10.1098/rstb.2011.0355

Stefanics, G., Haden, G. P., Sziller, I., Balazs, L., Beke, A., & Winkler, I. (2009). Newborn infants process pitch intervals. *Clin Neurophysiol, 120*(2), 304-308. doi: 10.1016/j.clinph.2008.11.020

Stringer, S. M., Perry, G., Rolls, E. T., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol Cybern, 94*(2), 128-142. doi: 10.1007/s00422-005-0030-z

Tew, S., Fujioka, T., He, C., & Trainor, L. (2009). Neural representation of transposed melody in infants at 6 months of age. *Ann N Y Acad Sci, 1169*, 287-290. doi: 10.1111/j.1749-6632.2009.04845.x

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn Sci, 13*(12), 532-540. doi: 10.1016/j.tics.2009.09.003

Chapter 44

Neural Correlates of Auditory-Tactile Integration in Musical Meter Perception

Juan Huang[1,2], Tianxu Wu[3], Steven Hsiao[1], and Xiaoqin Wang[2,3]

[1] *Zanvyl Krieger Mind/Brain Institute, The Johns Hopkins University, Baltimore, MD, USA*
[2] *Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, MD, USA*
[3] *Tsinghua-Johns Hopkins Joint Center for Biomedical Engineering Research and Department of Biomedical Engineering, Tsinghua University, Beijing, China*

Contact author: JH, jhuang7@jhu.edu

Abstract

Previously we reported that auditory and tactile inputs are seamlessly integrated to form the percept of meter in human subjects (Huang, et al. 2012). In this study, we examined the neural correlates of auditory and tactile integration in meter perception. Event Related Potentials (ERP) was recorded from 15 human subjects while they were presented with unimodal auditory, unimodal tactile or bimodal sequences. We report that the ERP signals corresponding to missing notes are similar to those induced by both the metrically important and unimportant auditory or tactile notes depending on the testing conditions. The results demonstrate that ERP can be induced without actual sensory input (either auditory or tactile) as long as the "virtual" notes are embedded in a sequence with a valid meter structure. Furthermore, the ERP components corresponding to bimodal sequences are determined by the modality of the metrically important notes, but not the physical sensory properties of the notes. The shape of ERP induced by tactile notes in a sequence could be auditory ERP as long as the metrical important notes were auditory stimulation, and vice versa. These results demonstrate neural correlates of auditory and tactile integration of meter perception. Our findings further suggest that either auditory or tactile modality can play a dominant role in the processing of meter information.

Keywords: Musical Meter, Tactile, Auditory, Multisensory Integration, Electroencephalography (EEG), Event Related Potentials (ERP)

1 Introduction

Meter perception is fundamental to the perception of music, which is determined by the temporal pattern of musical notes, such as periodicity, rhythm, and tempo (London, 2004; Phillips-Silver et al., 2010). Previous studies have explored the neural mechanisms of the perception of musical meter. In one line of studies, by recording the electroencephalogram (EEG) signals while subjects listening to music, researchers found steady state-evoked potentials (SS-EPs) being elicited at frequencies corresponding to beat in music pieces (Nozaradan et al., 2011). Furthermore, the periodicity of EEG responses seem to relate to the perception of rhythm rather than the actual acoustic energy of music piece (Nozaradan et al., 2012). Previous studies have also observed that event-related potentials (ERP) corresponding to beat or meter cycle induced at different time points are different in amplitude (Iversen et al., 2009; Fujioka et al.,

2010; Schaefer et al., 2011). However, whether the entrainment of neuronal resonance generated SS-EPs at the beat frequency or the superimposition of transient ERP reflects the perception of beat or meter remains unknown.

In a natural environment, it is common that an event occurring often associates with multiple sensory representations. For example, acoustic ringing of a wind-bell occurs together with the visual swinging pattern of the bell as well as the somatosensory signal caused by the variation of the wind power, which results in similar temporal patterns in acoustic, visual, and somatosensory stimulation. Identifying same or similar temporal patterns across sensory modalities is crucial for perceptually grouping different sensory inputs into a single event, and temporal pattern recognition across multiple sensory modalities become a key issue of multisensory integration. Earlier studies have focused on determining multisensory temporal relations, such as the perception of simultaneity or correspondence over discrete pairs of different sensory inputs (Spence and Squire, 2003; Stone et al., 2001; Recanzone, 2003; Fujisaki et al., 2004; Watanabe and Shimojo, 2001; van Eijk et al., 2008), discrimination the synchrony between stimuli (McGurk and MacDonald, 1976; Meredith et al., 1987; Shams et al., 2000), examination of the time window where the perception of one sensory input being influenced by another sensory input (Hirsh and Sherrick, 1961; Kristofferson, 1967; Zampini et al., 2005; Vroomen and Keetels, 2010; Denison et al., 2013), or multisensory integration between matching or unrelated temporal patterns across sensory modalities (Parise et al., 2012). Imaging studies have shown that comparing with unrelated temporal pattern, coincident temporal pattern across multisensory inputs produce larger BOLD responses in multisensory and primary sensory cortical areas (Noesselt et al., 2007). This line of studies mostly focused on the synchronous and/or the common temporal structure between two sensory events. However, sensory grouping could occur across different sensory modalities, in which a temporal pattern is formed by events from multiple sensory channels while no pattern structure is present within any of the individual sensory sequence, is less studied.

In a recent psychophysics study, we reported that auditory and tactile inputs were seamlessly integrated to form the percept of meter in human subjects who perceptually extracted music meter pattern from sequentially presented auditory and tactile inputs (Huang, et al. 2012). This study demonstrated that sensory grouping happens across auditory and tactile modalities. But the neural mechanisms underlying such cross modal sensory grouping remains unclear. In the present study, we examined the neural correlates of auditory and tactile integration in meter perception in human subjects using EEG recording and ERP analysis.

2 Methods

We recorded ERP from 15 human subjects while they were presented with unimodal auditory, unimodal tactile, or bimodal auditory and tactile note sequences. Stimuli were sequences consisting of 24 notes that were each 525 ms in duration. Fifteen of the temporal units were sine wave notes and nine were silent units. Each sine wave notes consisted of a 125 ms 180 Hz pure tone (for auditory modality) or a 180 Hz vibration (for tactile modality) followed by a 400 ms silent period. Each silent unit consisted of 525 ms of silence. The temporal patterns of the note sequences were adapted from our previous psychophysics study (Huang et al., 2012) and could be categorized as triple meter or duple meter. Each triple sequence had 8 measures with 3 units in each measure, and a duple sequence had 6 measures with 4 units in each measure. The first unit of each measure was always a sine wave note that was 3 dB higher in intensity than the rest notes in the sequence, which was called “accented notes” (AN). The rest of sine wave notes were called “unaccented notes” (UN). The 9 silent units were called “missing notes” (MN). Following this rule, a triple sequence had 8 AN, 7 UN, and 9 MN notes whereas a duple sequence had 6 AN, 9 UN, and 9 MN notes. Auditory stimulation was presented to the left ear of participants through an inserted headphone, and tactile stimulation was presented through a tactile vibration stimulator placed on subjects’ left index finger tip. For control condition, AN and UN notes for

unimodal sequences and auditory/tactile notes for bimodal sequences were presented in a randomized order, so that no meter structure was formed. In order to keep subjects' attention focused on the stimulation, the last note of 10% of the sequences was significantly increased in both amplitude and frequency. Subjects were also asked to detect an odd note at the end of each note sequence by pressing a button.

EEG data were recorded from fifteen subjects using a 64 channel Neuroscan system (Neuroscan Inc.) with a sampling rate of 1000 Hz. Electrodes were placed according to the 10–10 system and the reference electrode was located at the nose tip. Four recording sessions were conducted for each subject: uni-modal auditory, uni-modal tactile, bi-modal with auditory AN and tactile AN notes, and bi-modal with tactile AN and auditory UN notes. The four recording sessions were conducted to participants in a randomized order.

3 Results

EEG epochs were obtained by segmenting the data from -1 to 12 s relative to the onset of each note sequences (first note of the sequence). The epochs with abnormal variances (out of the 97% confidence interval) were rejected to avoid irregular EEG response. According to the note type (accented note, unaccented note or missing note) of each sequences, EEG epochs of each single note were abstained by segmenting epochs of each sequences from -0.08 to 0.5 s relative to the onset of each note. ERP corresponding to AN, UN, and MN notes were then averaged across sequence trials.

Data show that for metrical sequences, MN notes generated significant ERP components were similar to what AN and UN notes generated. However, for random sequences, MN notes did not induce significant ERP components. In addition, MN notes related ERP components did not develop during the first 1 or 2 measures of metrical sequences, but occurred obviously in the last few measures. The results of EEG components to MN notes demonstrate that ERP signals can be induced without actual sensory input (either, auditory or tactile) as long as the "virtual" notes are embedded in a sequence with a meter structure. Meanwhile, MN notes related ERP components observed only in the last few measures of metrical sequences indicate that the MN generated ERPs reflect the perception of meter rather than the perception of rhythm.

Data also shown that the shape and latency of ERP components corresponding to bimodal sequences with metric patterns was largely determined by the modality of the metrically important notes, e.g. the modality of AN notes. When the AN notes were auditory stimulation, the overall ERP waveforms corresponding to AN, UN, and MN notes were all similar to those induced by unimodal auditory notes; whereas the ERP waveforms corresponding to bimodal sequences with tactile stimulation as the AN notes were similar to those induced by the unimodal tactile notes. These results indicate that ERP components are not simply sensory modality dependent, but reflect an integrated form across auditory, tactile, and missing notes. The sensory modality independent ERP could be the neural correlates of auditory and tactile integration of meter perception. Such perceptual grouping cross modalities was closely linked to the cognitive aspects of perception but less to the physical properties of sensory inputs. These results indicate that the brain treats inputs from different sensory modalities as uniformed events, as long as the inputs from different sensory modalities form a regular temporal pattern or structure, which also suggests that multisensory integration was a global perceptual phenomenon that superimposes local processing of individual sensory events.

4 Discussion

Findings of this study show that the ERP components corresponding to missing notes signal that the brain makes a prediction of an event to occur based on the regularity of the presented note sequence. The perception of musical meter from note sequence containing missing notes relies

on the ability of the brain to extract temporal pattern from the sequence. The fact that missing notes in first few measures did not produce reliable ERP components suggests that the regularities of the note sequence was not extracted or perceived till a subject's brain had the chance to sample part of the sequence to estimate the regularity. As more notes of the sequence being presented, the cues that the brain uses to determine the regularity of the sequence become sufficient. Thus the missing notes of the rest of the sequence being treated by the brain as temporal units consisting of a regular temporal pattern, producing ERP responses similar to that accented and un-accented notes generated. In our study, the ERP components corresponding to MN notes observed in bimodal conditions occurred and its shape and latencies were similar to the ERP components corresponding to metrically important notes. In comparison, MN notes in random metrical sequences did not generate ERP components. In such cases, again, the brain treats a silent unit in bimodal metric sequences the same as an AN note, reflecting the prediction of an event occurring in the temporal structure.

Another line of studies exploring the role of attention has shown that attention enforces phase resetting and augments the neuronal oscillations to the attended stimulus stream (Lakatos et al., 2013). The attention produced neural entrainment corresponds to the detection of a rhythm in multisensory tasks in this case. However, whether the attention directed to one sensory stream or the rhythm embedded in the sensory stream is the critical factor to enhance the neural entrainment and reset the phase is not clear. In the bimodal sequence condition of our study, subjects' attention was directed to "an odd note or vibration at the end of the sequence", which was away from specific sensory modality and the rhythm of either auditory or tactile streams since neither streams contained any meter pattern. The missing notes related ERP was not caused by attention, but the perception of the meter. The results show a significant ERP complex corresponding to missing notes fitting in the EEG response wave form, meaning the brain successfully predicts the rhythm pattern based on the perceptual integration of inputs from both auditory and tactile channels.

References

Denison R.N., Driver J., Ruff C.C. (2013) Frontiers in Psychology: Perception Science 3, Article 619. doi: 10.3389/fpsyg.2012.00619

Fujioka T., Zendel B.R., Ross B. (2010) Endogenous neuromagnetic activity for mental hierarchy of timing. J Neurosci. 30: 3458-3466

Fujisaki W., Shimojo S., Kashino M., and Nishida S. (2004) Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778

Hirsh I.J., and Sherrick C.E. (1961) Perceived order in different sense modalities. *J. Exp. Psychol.* 62, 423–432

Huang J., Gamble D., Sarnlertsophon K., Wang X.Q., Hsiao S. (2012) Feeling Music: Integration of Auditory and Tactile Inputs in Musical Meter Perception. *PLoS One,* Doi: 10.1371/ journal.pone.0048496. Epub 2012 Oct 31.

Iversen J.R., Repp B.H., Patel A.D. (2009) Top-down control of rhythm perception modulates early auditory responses. Ann NY Acad Sci 1169: 58-73

Kristofferson A.B. (1967) Successiveness discrimination as a two state, quantal process. *Science* 158, 1337–1339

Lakatos P., Musacchia G., O'Connel M.N., Falchier A.Y., Javitt D.C., and Schroeder C.E. (2013) The spectrotemporal filter mechanism of auditory selective attention. Neuron 77: 750-761

London J. (2004) Hearing in time: psychological aspects of musical meter. New York: Oxford UP

McGurk H., and MacDonald J. (1976) Hearing lips and seeing voices. *Nature* 264, 746–748

Meredith M.A., Nemitz J.W., and Stein B.E.(1987) Determinants of multisensory integration in superior colliculus neurons. I.Temporal factors. *J. Neurosci.* 7, 3215–3229

Noesselt T., Rieger J.W., Schoenfeld M.A., Kanowski, M., Hinrichs, H., Heinze, H.J., et al. (2007) Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441

Nozaradan S., Peretz I., Missal M., Mouraux A. (2011) Tagging the neuronal entrainment to beat and meter. J Neurosci 31:10234 –10240

Nozaradan S., Peretz I., Mouraux A. (2012) Selective Neuronal Entrainment to the Beat and Meter Embedded in a Musical Rhythm. J. Neurosci 32(49): 17572-17581

Nozaradan S., Peretz I., Mouraux A. (2012) Steady-state evoked potentials as an index of multisensory temporal binding. Neuroimage 60:21–28

Parise C.V., Spence C., and Ernst M.O. (2012) When correlation implies causation in multisensory integration. *Curr. Biol.* 22, 46–49

Phillips-Silver J., Aktipis C.A., Bryant G.A. (2010) The ecology of entrainment: foundations of coordinated rhythmic movement. Music Percept 28:3–14

Recanzone G.H. (2003) Auditory influences on visual temporal rate perception. *J. Neurophysiol.* 89, 1078–1093

Schaefer R.S., Vlek R.J., Desain P. (2011) Decomposing rhythm processing: electroencephalography of perceived and self-imposed rhythmic patterns. Psychol. Res. 75: 95-106

Shams L., Kamitani Y., and Shimojo S. (2000) What you see is what you hear. *Nature* 408, 788

Spence C., and Squire S. (2003) Multisensory integration: maintaining the perception of synchrony. *Curr. Biol.* 13, R519–R521

Stone J.V., Hunkin N.M., Porrill J., Wood R., Keeler V., Beanland M., etal. (2001) When is now? Perception of simultaneity. *Proc. Biol. Sci.* 268, 31–38

Van Eijk, R., Kohlrausch, A., Juola, J., and van dePar, S. (2008) Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. *Percept. Psychophys.* 70, 955–968

Vroomen J., and Keetels M. (2010) Perception of inter sensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884

Watanabe K., and Shimojo S. (2001) When sound affects vision: effects of auditory grouping on visual motion perception. *Psychol. Sci.* 12, 109–116

Zampini M., Guest S., Shore D.I., and Spence C. (2005) Audiovisual simultaneity judgments. *Percept. Psychophys.* 67, 531–544

Chapter 45

Do Zwicker Tones Evoke a Musical Pitch?

Hedwig E. Gockel, Robert P. Carlyon
MRC Cognition and Brain Sciences Unit, Cambridge, UK

Contact author: H. E. G., hedwig.gockel@mrc-cbu.cam.ac.uk

Abstract

It has been argued that musical pitch, i.e. pitch in its strictest sense, requires phase locking at the level of the auditory nerve. The aim of the present study was to assess whether a musical pitch can be heard in the absence of peripheral phase locking, using Zwicker tones (ZTs). A ZT is a faint, decaying tonal percept that arises after listening to a band-stop (notched) broadband noise. The pitch is within the frequency range of the notch. Several findings indicate that ZTs are unlikely to be produced mechanically at the level of the cochlea and, therefore, there is unlikely to be phase locking to ZTs in the auditory periphery. In stage I of the experiment, musically trained subjects adjusted the frequency, level, and decay time of an exponentially decaying sinusoid so that it sounded similar to the ZT they perceived following a broadband noise, for various notch positions. In stage II, subjects adjusted the frequency of a sinusoid so that its pitch was a specified musical interval below that of either a preceding ZT or a preceding sinusoid (as determined in stage I). Subjects selected appropriate frequency ratios for ZTs, although the standard deviations of the adjustments were larger for the ZTs than for the equally salient sinusoids by a factor of 1.1-2.2. The results suggest that a musical pitch may exist in the absence of peripheral phase locking.

Keywords: auditory afterimage, tonal percept, band-stop noise, notched noise, phase locking, pitch match, frequency adjustment, pitch salience, musical interval

1 Introduction

Periodic sounds produce periodic patterns of phase-locked activity in the auditory nerve. It has been argued that this temporal code is the basis for our sensation of pitch, and, specifically, that musical pitch i.e. pitch in its strictest sense, requires phase locking. In this vein, most current models of pitch perception rely on the precise timing of action potentials, or spikes, within the auditory nerve (Cariani & Delgutte, 1996; Meddis & O'Mard, 1997). Furthermore, the precision of phase locking weakens at high frequencies, and the putative upper limit of phase locking at frequencies of about 4-5 kHz (Johnson, 1980; Palmer & Russell, 1986) has been used to explain the finding that the ability to recognize melodies, the accuracy of musical interval judgements, and frequency discrimination are all severely degraded for pure tones above about 4-5 kHz (Ward, 1954; Attneave & Olson, 1971; Moore, 1973; Sek & Moore, 1995).

There exist, however, some observations that cast doubt on the generally accepted assumption that phase locking is necessary for musical pitch. For example, while most subjects in Ward's (1954) study were unable to make octave judgements when the reference frequency, f1, was above 2700 Hz, for two of his subjects the variability in octave judgements was essentially the same for f1=5 kHz (with the octave match at ~10 kHz) as for lower frequencies. However, octave judgements were more difficult and subjects took a greater time when f1 was 5 kHz. Ward

suggested that experience might play a role, as these two subjects were the only ones with experience in judging the pitch of pure tones. Similarly, Burns and Feth (1983) asked musically trained subjects to adjust various musical intervals for reference frequencies of 1 kHz and 10 kHz. Even for the high frequency, all three subjects could do the task. For the 10-kHz reference, the standard deviations, SDs, averaged across all musical intervals, were about 3.5-5.5 times larger than for the 1-kHz reference. This increase was less than that observed for unison adjustments, which were taken as an estimate of the difference limen for frequency, DLF. Burns and Feth concluded that their results were not incompatible with a temporal basis for both frequency discrimination and musical interval adjustment, as phase locking information decreases with increasing frequency. More recently, Oxenham et al. (2011) showed that complex tones whose audible harmonics all fall above 6 kHz can evoke a robust sense of pitch and musical interval, but high-frequency pure tones did not, even though the just noticeable difference for the latter was less than 1.5%. They concluded either that there is sufficient phase-locking present at high frequencies to derive complex musical pitch, or that a complex pitch can be derived in the absence of phase locking. Pertaining to this, Moore and Ernst (2012) reported DLFs for center frequencies from 2 to 14 kHz that were consistent with the idea that there is a transition from a temporal to a place mechanism at about 8 kHz, rather than at 4-5 kHz, as commonly assumed.

Here we assessed whether a musical pitch can be heard in the *absence* of phase locking, using Zwicker tones, ZTs (Zwicker, 1964). A ZT is a faint, decaying tonal percept that can arise following the presentation of a notched broadband noise and that can last up to 5-6 s. The pitch is always within the frequency range of the notch, and depends on the level of the noise and the width of the notch. Several findings indicate that ZTs are unlikely to be produced mechanically at the level of the cochlea. Normally, no otoacoustic emissions (OAEs) were found at the frequency corresponding to the pitch of the ZT, except in the rare case when a subject had a spontaneous OAE which could be made temporarily audible by a preceding notched noise (without increasing its physical level), and no beating was observed between ZTs and a soft physical tone (Krump, 1993). Also, while low frequency tones at moderate to high levels can affect the level of evoked OAEs, they did not affect the ZT (Wiegrebe et al., 1995). Thus, no mechanical activity correlated to the ZT exists in the cochlea, and there is unlikely to be phase locking to ZTs in the auditory periphery. The question addressed here is whether ZTs can evoke a musical pitch.

2 Methods

2.1 Experimental Design

Four young normal-hearing musically trained subjects took part. In stage I, subjects adjusted the frequency, level, and decay time of an exponentially decaying sinusoid so that it sounded similar to the ZT they perceived following a broadband noise, for various notch positions. In stage II, subjects adjusted the frequency (and level) of a sinusoid so that its pitch was a specified musical interval below that of either a preceding ZT or a preceding sinusoid (and to be equally loud). Importantly, for each subject, the reference sinusoids corresponded to those that were adjusted in stage I to sound similar to, i.e. have equal pitch, loudness and decay time as, the ZTs. The precision of the musical interval adjustments for the ZTs and the matched sinusoids (PT condition) was compared.

2.2 Stimuli and General Procedure

To evoke ZTs, 5-s (including 20-ms onset and offset ramps) notched broadband noises (30-16000 Hz) were presented diotically at an rms level of 51 dB SPL. From one match to the next, the lower edge frequency (LEF) of the notch could take one of eight values: 2000, 2144, 2297, 2460, 2633, 2818, 3014, and 3500 Hz. The higher edge frequency (HEF) of the notch was always

1.5 times the LEF. The adjustable sinusoid had a maximum duration of 5 s (including 20-ms onset and offset ramps) and followed the ZT-exciting noise with an inter-stimulus interval of 5.5 s. After cessation of the sinusoid, the subject indicated by button presses the desired direction of change for the frequency, level, and (stage I only) time constant of the sinusoid for the next presentation, and/or initiated the next sound presentation, i.e. the next trial. In each trial, the subject was allowed an unlimited number of button presses before s/he initiated the next trial. For each parameter that needed adjusting, three different step sizes were available for each direction of change. The smallest available step sizes were 1/16 semitone, 1 dB, and a factor of $2^{1/4}$ for the frequency, level, and time constant adjustments, respectively. Subjects could take as many trials as they liked to finish a match. The first session (2 hrs including breaks) was considered practice, and matches from this session were discarded. For the experiment proper, typically about 10 matches were collected for each condition from each subject.

In stage I, in most cases, the adjusted frequency of the matched sinusoid increased with increase in LEF and was a factor of 1.1-1.2 above the LEF. Two of the subjects reliably matched higher frequencies (factor of 1.3-1.4 above the LEF) in two or three LEF conditions, leading to functions where the matched frequency sometimes did not increase with increase in LEF. For the present purposes, these cases were excluded from further analysis of stage II data, as they could be due to spontaneous OAEs that became audible after listening to the notched noise (Krump, 1993; Wiegrebe et al., 1995). The initial level and time constant of the matched sinusoids had mean values that ranged from 5-23 dB above threshold in quiet and from 1.3-5.1 s, respectively, across subjects and LEFs. In the following, we consider the interval adjustments of stage II of the experiment only.

2.3 Musical Interval Adjustments

In stage II, subjects adjusted a sinusoid to be a specified musical interval below the preceding reference tone. The musical intervals that had to be adjusted were a minor third (3 semitones down) and a perfect fifth (7 semitones down). In different sessions, the reference tone was either a ZT or a physically presented pure tone (PT). The latter corresponded to the matched tones from stage I. From one match to the next, the reference tone could be any one of the set of eight. The general trial structure was the same as in stage I. Typically about 12 (at least 10) interval adjustments were collected for each condition from each subject.

3 Results

Figures 1 and 2 show, for each subject, several measures of the accuracy, repeatability, and ease of the musical interval adjustments in the ZT and PT conditions. Figure 1 shows the geometric mean (and SDs across the 8 reference frequencies or LEFs) of the ratio of the adjusted frequency of the variable pure tone to the expected frequency. The expected frequency was determined on the equal temperament scale. That is, the expected frequencies for the minor third and the perfect fifth were exactly 3 semitones (a factor of 1/1.189) and 7 semitones (a factor of 1/1.498) below the reference frequency. For all subjects, the adjusted frequencies were somewhat flat, i.e. slightly lower than expected, leading to somewhat larger musical intervals. However, this was true for both the PT and the ZT conditions (different coloured bars). Generally, the adjusted frequencies were within about 2% of the expected frequencies, with no systematic differences between the ZT and the PT conditions. Thus, subjects were able to match musical intervals with good accuracy for both conditions.

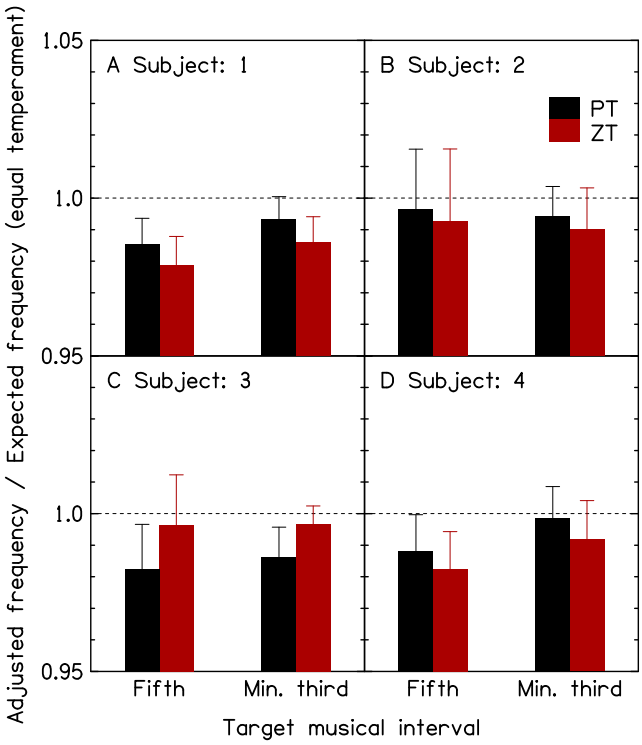


Figure 1. Geometric mean (and SDs across the 8 reference frequencies or LEFs) of the ratio of the adjusted frequency of the variable pure tone to the expected frequency for musical intervals of a perfect fifth and a minor third.

To compare the reliability of the musical interval adjustments in the ZT conditions with that in the matched PT conditions, we calculated the ratio of the respective SDs. This was done as follows: (1) Separately for each subject, each musical interval (minor third, perfect fifth), each reference condition (ZT, PT) and each LEF (or reference PT tone frequency) we determined the SD of the adjusted frequencies across the 12 interval adjustments expressed as a percentage of the geometric mean; (2) These were geometrically averaged across the eight LEFs (or PT frequencies) to obtain the representative SD; (3) The representative SD for the ZT was divided by the representative SD for the PT. This ratio is shown in the group of two bars on the left-hand side of each panel in Figure 2. Musical interval adjustments for the ZTs were somewhat more variable than for the matched PTs; the ratio of the SDs was between 1.1 and 2.2. Four univariate analyses of variance conducted separately on the data for each subject (using the SD of the logarithms of the individual adjusted frequencies within a given LEF or reference frequency as input data) showed that for three of the subjects the SDs were significantly larger for the ZT than for the PT conditions (with $p<0.01$ for subjects 1 and 2, and $p<0.05$ for subject 3). The number above each bar in Figure 2 indicates the size of the representative SD for the PT condition in percent (ranging from 0.9%-1.5%). The representative SD for the ZT condition ranged from 1.2%-2.5%. By comparison, the SDs observed by Burns and Feth (1983) for the same musical interval adjustments were on average 1.2% and 4.2% for 1-kHz and 10-kHz 70-dB SPL reference tones, respectively.

The reliability of adjustments can be increased by listening more often to the stimuli, and thus needs to be considered as well. The group of two bars on the right-hand side of each panel in Figure 2 shows the ratio of the average number of listening times (trials) in the ZT and the PT conditions. The number above each bar indicates the average number of trials (n-listen) for the PT condition. There were inter-individual differences in how many trials a subject needed on average to make a match. For example, in the PT conditions, subject 1 needed on average 7.6 trials while subject 3 needed on average 5.7 trials. However, n-listen was very similar for the two conditions, i.e. the ratio was very close to one. If anything, n-listen was very slightly lower

for the ZTs than for the PT. This means that the achieved accuracy and reliability in the ZT conditions did not come at the cost of listening more often in the ZT than the PT conditions.

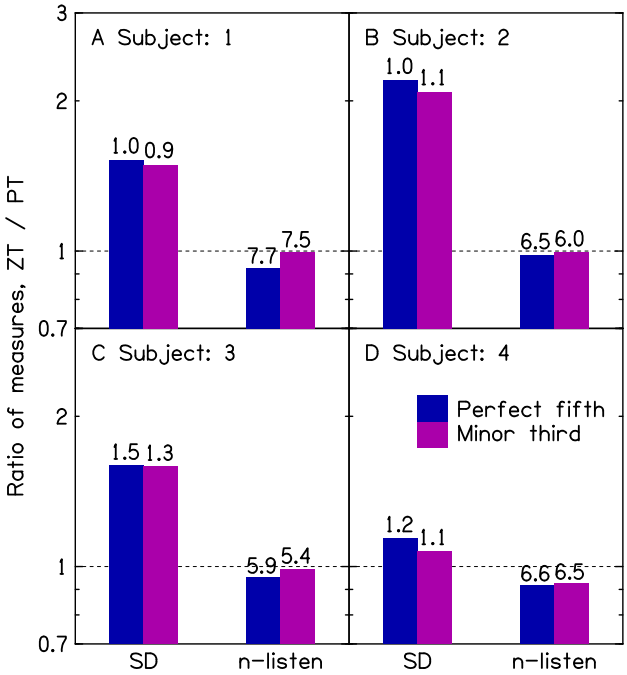


Figure 2. Ratio of the geometric mean (across the 8 reference frequencies or LEFs) of the SDs (left-hand side) and of the average number of trials (right-hand side) between ZT and PT conditions, for musical intervals of a perfect fifth and a minor third. The value of each measure for the PT condition is given by the number above the corresponding bar; for the SD this is the representative SD (see text), expressed as a percentage of the geometric mean of the adjusted frequencies.

4 Discussion

The results showed that, on average, subjects selected similar frequencies in a musical interval adjustment task irrespective of whether the reference tone was a ZT or an equally salient pure tone. The adjusted frequencies were slightly flat in both cases. This might partly result from subjects' bias towards a "just scale" with musical intervals corresponding to integer ratios of 6/5 and 3/2 for the minor third and the perfect fifth, respectively. If this was the only underlying reason, the "bias" should have been larger for the minor third than for the perfect fifth, which was not observed. The SDs of the musical interval adjustments were only slightly larger (a factor of 1.1-2.2) for the ZTs than for the PTs, with no increase in listening time. Thus, overall the results suggest that a weak musical pitch can exist in the absence of peripheral phase locking.

Burns and Feth (1983) had three musically trained subjects adjust various musical intervals for reference frequencies of 1 kHz and 10 kHz. For the same musical intervals as used in the present study, SDs in their data were a factor of 1.6-11.7 (mean of 4) times larger for the 10-kHz reference than for the 1-kHz reference. Thus, the increase in SDs for the 10-kHz frequency was larger than that observed here for the ZTs. Burns and Feth (1983) interpreted their results in terms of phase locking, and a decrease thereof with increasing frequency. The present results suggest that their relatively large increase in SDs at 10 kHz reflects the combined effect of lack of familiarity with high-frequency pure tones and the transition from a temporal code to a place code.

Acknowledgments

Supported by intramural funding from the MRC. Thanks to Brian Moore for helpful comments.

References

Attneave, F., & Olson, R. K. (1971). Pitch as a medium - New approach to psychophysical scaling. *American Journal of Psychology,* 84, 147-166.

Burns, E. M., & Feth, L. L. (1983). Pitch of sinusoids and complex tones above 10 kHz. In R. Klinke & R. Hartmann (Eds.). *Hearing - Pysiological Bases and Pyschophysics* (pp. 327-333). Berlin: Springer.

Cariani, P. A., & Delgutte, B. (1996). Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. *Journal of Neurophysiology,* 76, 1717-1734.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *Journal of the Acoustical Society of America,* 68, 1115-1122.

Krump, G. (1993). Beschreibung des akustischen Nachtones mit Hilfe von Mithörschwellenmustern. Ph.D. thesis, Technical University München, Germany.

Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *Journal of the Acoustical Society of America,* 102, 1811-1820.

Moore, B. C. J. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America,* 54, 610-619.

Moore, B. C. J., & Ernst, S. M. A. (2012). Frequency difference limens at high frequencies: Evidence for a transition from a temporal to a place code. *Journal of the Acoustical Society of America,* 132, 1542-1547.

Oxenham, A. J., Micheyl, C., Keebler, M. V., Loper, A., & Santurette, S. (2011). Pitch perception beyond the traditional existence region of pitch. *Proceedings of the National Academy of Sciences of the United States of America,* 108, 7629-7634.

Palmer, A. R., & Russell, I. J. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research,* 24, 1-15.

Sek, A., & Moore, B. C. J. (1995). Frequency discrimination as a function of frequency, measured in several ways. *Journal of the Acoustical Society of America,* 97, 2479-2486.

Ward, W. D. (1954). Subjective musical pitch. *Journal of the Acoustical Society of America,* 26, 369-380.

Wiegrebe, L., Kössl, M., & Schmidt, S. (1995). Auditory sensitization during the perception of acoustical negative afterimages - Analogies to visual processing. *Naturwissenschaften,* 82, 387-389.

Zwicker, E. (1964). 'Negative afterimage' in hearing. *Journal of the Acoustical Society of America,* 36, 2413-2415.

Chapter 46

Speech Coding in the Midbrain: Effects of Sensorineural Hearing Loss

Laurel H. Carney

Departments of Biomedical Engineering, Neurobiology & Anatomy, Electrical & Computer Engineering, University of Rochester, Rochester, NY, USA

Duck O. Kim, Shigeyuki Kuwada

Department f Neuroscience, University of Connecticut Health Center, Farmington, CT, USA

Contact author: LHC, Laurel.Carney@Rochester.edu

Abstract

In response to voiced speech sounds, auditory-nerve (AN) fibres phase-lock to both harmonics near best frequency (BF) and to the fundamental frequency (F0) of voiced sounds. Due to nonlinearities in the healthy ear, phase-locking in each frequency channel tends to be dominated either by a single harmonic, for channels tuned near formants, or by F0, for channels between formants. The alternating dominance of these factors sets up a robust pattern in the F0-synchronized rate across best frequency (BF). This profile of a temporally coded measure is transformed into a mean rate profile at the level of the midbrain (inferior colliculus, IC), where neurons are sensitive to low-frequency fluctuations. In the impaired ear, the F0-synchronized rate profile is affected by several factors: Reduced synchrony capture decreases the dominance of a single harmonic near BF on the response, resulting in more widespread F0-synchrony. Elevated thresholds also reduce the effect of rate saturation, resulting in increased F0-synchrony. Wider peripheral tuning results in a wider-band envelope with reduced F0 amplitude. In general, sensorineural hearing loss reduces the *contrast* in AN F0-synchronized rates across BF. Here, we use computational models for AN and IC neurons to illustrate the F0-synchronized rate profiles set up in response to voiced speech sounds. The IC models include those for neurons with band-enhanced, band-suppressed, and hybrid modulation transfer functions. These models provide an illustration of the effects of peripheral hearing loss on central representations of voiced speech sounds.

Keywords: vowels, modulation transfer function, neural coding

1 Introduction

The impact of sensorineural hearing loss (SNHL) on speech communication is perhaps the most significant problem faced by listeners with hearing loss. A puzzle regarding the effects of SNHL is that relatively small amounts of loss can have a surprisingly large impact on listeners, especially in noisy acoustic environments. Here we focus on the effects of SNHL on the representation of voiced speech sounds in the auditory periphery and in the midbrain (inferior colliculus, IC). Responses of models for auditory-nerve (AN) fibres and IC neurons provide insight into how changes in the temporal structure of AN responses can have a significant impact at the level of

the IC. Neurons in the IC are sensitive to low-frequency rate fluctuations, such as rates that are synchronized to the envelopes of amplitude-modulated (AM) stimuli.

Vowels are an essential component of all languages, and play an important role in carrying information, especially in running speech (e.g. Kewley-Port et al., 2007). Vowel sounds are characterized by voicing, which results in a spectrum with a large peak at the fundamental frequency (F0), related to voice pitch, and peaks at the harmonics of F0. The harmonic spectrum is shaped by vocal tract resonances, creating wide spectral peaks referred to as formants. The locations of the formants, especially the lowest two frequency formants (F1 and F2), distinguish different vowel sounds (Fant, 1970). Neural coding of these spectral peaks is essential for understanding how the brain processes speech.

Most of the work on neural coding of vowels has focused on questions of rate-place and temporal coding of fine-structure of the formants at the level of the AN (e.g. Sachs & Young, 1979; Young & Sachs, 1979; Delgutte & Kiang, 1984; reviewed by Young, 2008). Consideration of midbrain responses to vowels shifts the focus to the peripheral representation of low-frequency fluctuations ($\lesssim$ 250 Hz) associated with F0. IC cells are particularly sensitive to amplitude modulation in the frequency range of F0 (Langner, 1992; Carney et al., 2015; Kim et al., 2015a,b). In the healthy ear, nonlinear response properties in the periphery result in a systematic pattern of neural fluctuations that are synchronized to F0 or to higher harmonics near AN characteristic frequencies. In particular, synchrony capture results in the dominance of a single harmonic near a spectral peak on the fine-structure of AN responses to voiced sounds (Delgutte & Kiang, 1984; Deng & Geisler, 1987). After trauma-induced SNHL, synchrony capture is reduced, and AN fibres respond to multiple harmonics (Miller et al., 1997; Sachs et al., 2002). These studies were focused on the effect of synchrony capture on the fine-structure coding of AN fibres; however, this mechanism also sets up a strong pattern of low-frequency fluctuations across the population of AN responses. In the healthy ear, fibres tuned near formant peaks are "captured" by a single harmonic, and thus have responses with relatively flat envelopes. In contrast, fibres tuned to frequencies between formants respond to multiple harmonics, and thus have strong rate fluctuations at F0. The amplitude of F0-related neural fluctuations provides a robust code for the formant frequencies that is accentuated in the midbrain. In the ear with SNHL, the decrease in synchrony capture results in fibres that respond to multiple harmonics, thus F0-related fluctuations prevail across all driven frequency channels at the level of the AN. As a result, the profile of F0-related fluctuations that codes the formants, and that ultimately drives modulation-sensitive midbrain neurons, is diminished by SNHL.

In this study, we use computational models to illustrate vowel responses of populations of AN fibres with and without SNHL. We then use models for three types of modulation sensitivity in the IC to illustrate the effect of SNHL on population responses at the level of the midbrain.

2 Methods

The computational model used here for AN responses was Zilany et al. (2014). This model has a provision for including impairment due to outer and inner hair cell dysfunction (Zilany & Bruce, 2007), which reduces synchrony capture in a manner comparable to the loss observed in physiological studies of AN response in ears that have been acoustically traumatized (Miller et al., 1997). In the results here, SNHL was simulated by setting AN model parameters C_{OHC} and C_{IHC} to 0.3; these values simulate mild hearing loss, with threshold elevations ranging from about 15 dB at 500 Hz to 30 dB at 2000 Hz and higher.

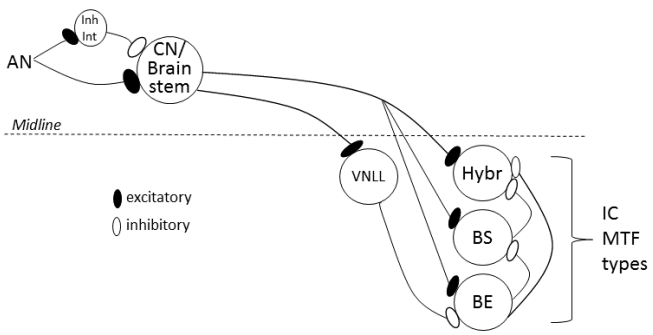
Three phenomenological models were used for IC cells to simulate the three types of rate modulation transfer functions (MTFs) that comprise the bulk of IC neurons (Kim et al., 2015a): band-enhanced, band-suppressed, and hybrid (Fig. 1). The first two MTF types have discharge rates in response to AM stimuli that are enhanced or suppressed relative to the response to unmodulated

stimuli (Kim et al., 2015a,b). In hybrid MTFs, the discharge rate is enhanced at some modulation frequencies and suppressed at other frequencies (Krishna & Semple, 2000; Kim et al., 2015a).

The band-enhanced model (Fig. 1, BE) was based on Nelson and Carney's (2004) same-frequency inhibitory-excitatory (SFIE) model. The key mechanism in this model is a dynamic interaction between relatively short-duration excitatory and relatively long-duration, delayed inhibitory synaptic potentials. The durations of the excitatory and inhibitory potentials, as well as the delay of the inhibition with respect to the excitation, determine the best modulation frequency (BMF) of the band-enhanced model cell.

An extension of the SFIE model was used to obtain the band-suppressed model (Fig. 1, BS). This model receives the same ascending excitatory input as the band-enhanced model, but it is inhibited by the band-enhanced model (Carney et al., 2015; Kim et al., 2015b). The BMF of the band-enhanced neuron that provides the inhibitory input determines the worst modulation frequency (WMF) for the band-suppressed model cell.

Finally, the hybrid MTF model receives the same ascending excitatory input as the other model IC neurons, but it is inhibited by both band-enhanced and band-suppressed cells (Fig. 1, Hybr). The relation between the BMF and WMF of the inhibitory inputs determines the shape of the hybrid model cell's MTF.



Synapse	Excitatory Tau (ms), Ampl	Inhibitory Tau (ms), Ampl, Delay (ms)
CN inputs	0.5, 1	2.0, 0.6, 1.0
BE inputs	BE cell: 1.0, 1 BE to BS cell: 2.0, 1	BE cell: 1.0, 1.8, 2.0 BE to BS cell: 6.0, 1.1, 2.0
BS inputs	2.0, 0.8	6.0, 7.0, 2.0
Hybrid inputs	2.0, 0.8	6.0, 8.0 (from BE) & 1.3 (from BS), 2.0

Figure 1. Schematic diagram of the extended SFIE model for IC neurons with three types of MTFs: band-enhanced (BE), band-suppressed (BS) and hybrid. For simplicity, the simulations presented here were driven only by contralateral inputs. An AN response excites and inhibits (via an interneuron) a model cell representing a cochlear nucleus (CN) or other brainstem-level neuron. The brainstem neuron excites all three IC cell types, and inhibits (via the ventral nucleus of the lateral lemniscus, VNLL) the BE model. The BE model inhibits the BS model (Carney et al., 2015; Kim et al., 2015b), and both BE and BS models provide inhibition to the hybrid model (Kim et al., 2015a). In each model, inputs were convolved with the post-synaptic potential waveforms, summed, and half-wave rectified to represent discharge probability as a function of time (Nelson and Carney, 2004).

3 Results

The responses of AN fibers to the vowel/æ/ (in "had") are illustrated in Fig. 2, which also shows a portion of the vowel waveform (Fig. 2a) and its spectrum (Fig. 2b). Discharge rate as a function of BF is shown in the left-hand panels for models of healthy (Fig. 2c) and impaired (Fig. 2e) populations. At conversational speech levels, these profiles encode the formant frequencies. However, in background noise, or as sound level increases, the profiles tend to saturate (though intelligibility does not decrease until high SPLs are reached (Studebaker et al., 1999)).

Figure 2 (d, f) illustrates several features of the temporal responses of AN fibers to vowels. These plots show the dominant components of the AN responses, which are the largest spectral components of the post-stimulus time histograms (Delgutte & Kiang, 1984; Miller et al., 1997;

Sachs et al., 2002). Dominant components for the healthy AN model (Fig. 2d) show synchrony capture, or dominance of the temporal responses by a single harmonic near the formant peak. The synchrony capture extends across a span of fibres with BFs near each formant frequency (red arrows). The dominance of these responses by a harmonic near the formants (orange arrows) results in absence of synchrony to F0 (black arrow) for BFs near formants. In contrast, fibres tuned between the formants are not dominated by a single harmonic. Because these fibres respond to multiple harmonics near their BFs (green arrow), their responses synchronize to the "beats" at the difference frequency between harmonics, which is F0 (black arrow).

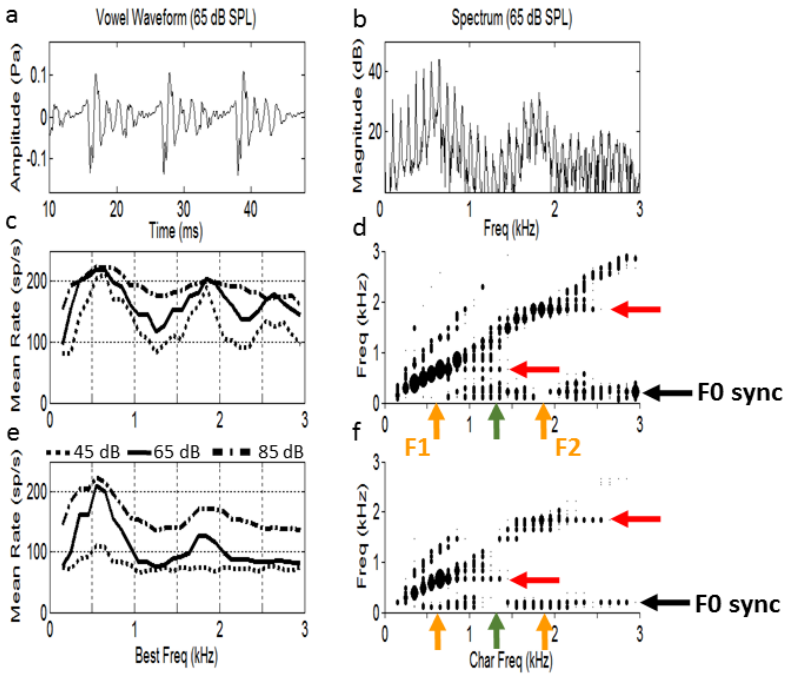


Figure 2. Time waveform (a) and spectrum (b) for the vowel /æ/ (in "had") at 65 dB SPL. The vowel signal was from Hillenbrand et al., 1995; F0=95 Hz. (c) Healthy and (e) impaired AN model rate profiles for 3 sound levels. (d,f) Dominant components of synchronized rate, found by taking the Fourier transform of the peri-stimulus time histogram, for the healthy and impaired responses at 65 dB SPL. Symbol sizes are related to the magnitude of the synchronized rate (components smaller than 15 sp/sec are not shown.)

The dominant components for model AN fibres with SNHL simulated by outer and inner hair cell dysfunction are shown in Fig. 2f. Physiological studies have shown that synchrony capture is weaker in these fibres (Miller et al., 1997). The model response illustrates reduced synchrony capture, especially to F2 (Fig. 2f, red arrow), even for the mild hearing loss in this simulation. The reduced synchrony capture is more evident in the pattern of synchrony to F0, which extends across all regions where there is sufficient energy to drive the neurons (Fig. 2f, black arrow). Thus, the profile of F0-synchronized rates differs qualitatively between the impaired and the healthy AN populations. For BFs tuned between F1 and F2 (green arrow), the response to F0 (black arrow) is reduced in the impaired nerve, whereas in the BF regions of F1 and of F2 (orange arrows), the response to F0 is increased in the impaired nerve. The net effect of SNHL is a reduction in the contrast of the F0-synchronized rates across BF.

Figure 3 shows examples of IC neural (left) and model (right) band-enhanced, band-suppressed, and hybrid MTFs. Kim et al. (2015a) found that, when tested with multiple types of AM envelopes (e.g., raised-sine envelopes with exponents of 1, 8 and 32), the bulk of IC neurons had one of these three types of rate MTFs. When a 25% change in rate with respect to the unmodulated response was the criterion for enhanced or suppressed rates, the band-suppressed type comprised nearly one half of the sample, and the band-enhanced and hybrid types each

comprised approximately one quarter. Model MTF shapes were determined by excitatory and inhibitory synaptic potential durations, amplitudes, and delays.

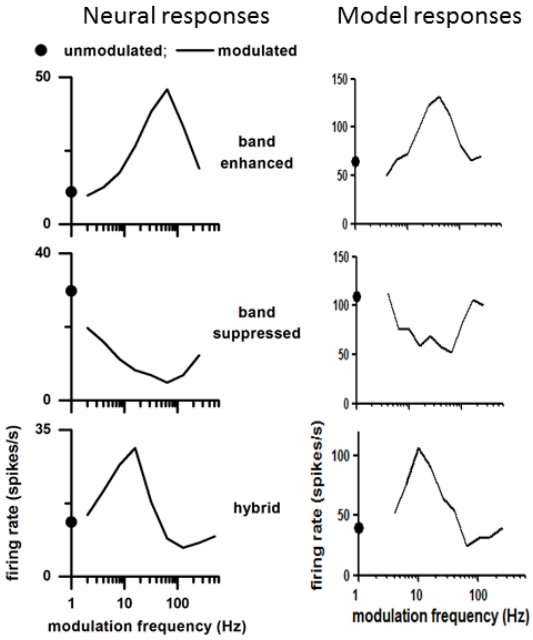


Figure 3. Left: Responses of IC neurons with the three types of rate modulation transfer functions (MTFs) that comprise the bulk of the IC neurons (from Kim et al., 2015a). Responses were recorded in the IC of awake rabbit. Right: Responses of three IC models, all with BF = 2 kHz. All neural and model responses are to 1-octave bandwidth sinusoidal AM noise (100% modulation depth), centered at BF. Filled dot is response rate to unmodulated noise.

Model responses to the vowel /æ/ for the three MTF types are shown in Fig. 4 for healthy and impaired model AN inputs. The healthy band-enhanced responses (Fig. 4a, blue) reflect the patterns of F0-synchronized rate shown in Fig. 2; responses are decreased near formant frequencies (orange arrows), because synchrony capture in the healthy AN reduces the low-frequency fluctuations that ultimately drive this IC type. Band-enhanced neurons tuned to frequencies between formants (Fig, 4a, green arrow) have strong responses due to the strong F0-synchronized rate in the healthy AN for these frequency channels (Fig. 2d, green arrow). The strong contrast in rate across frequency channels for the band-enhanced responses is robust across SPLs and in background noise (Carney et al., 2015).

The rate profile for healthy band-suppressed IC model neurons has peaks at the formant frequencies (Fig. 4a, magenta). These cells respond more strongly at frequencies where the low-frequency fluctuations are reduced (Fig. 2d, orange arrows). The peaks in the rate profile at the formant frequencies are very strong in the model hybrid population (Fig. 4a, green). This model result motivates further physiological and computational investigation of these cells. The hybrid neuron and model in Fig. 3 are most strongly driven by a contrast in neural fluctuations across modulation frequency (i.e., strong neural fluctuations in the 8-30 Hz range and reduced fluctuations in the 30-200 Hz range); other MTF shapes are also observed for hybrid neurons (Krishna and Semple, 2000).

Model IC responses with impaired AN inputs (Fig. 4b) differ dramatically from the healthy responses. The band-enhanced rate profile (Fig. 4b, blue) has peaks approximately where the healthy model has valleys, and vice versa. The impaired model profile is explained by reduced synchrony capture: the impaired AN model synchronizes to F0 whenever the fibres are adequately driven by stimulus energy (see Fig. 2f). Thus the impaired band-enhanced rate

profile reflects the energy spectrum of the vowel, unlike the healthy rate profile, which is inversely related to energy. The impaired band-suppressed (Fig. 4b, magenta) and hybrid (Fig. 4b, green) model responses also have peaks near F1 and F2. Unlike the healthy case, the inhibition from the shifted peaks in the impaired band-enhanced model suppresses the peaks in the band-suppressed and hybrid models. These results emphasize the fact that the strong rate profiles in the healthy IC model are not simply explained by the rate vs.BF profile of the AN. The healthy band-suppressed and hybrid profiles are created by a synchrony-to-rate transformation between the AN and IC and enhanced by disinhibition at the formant frequencies. In general, the rate profiles for all three response types are qualitatively different for this mildly impaired model, in a manner that can be explained based on the differences in the F0-synchronized rate between the healthy and impaired AN model responses (Fig. 2d,f).

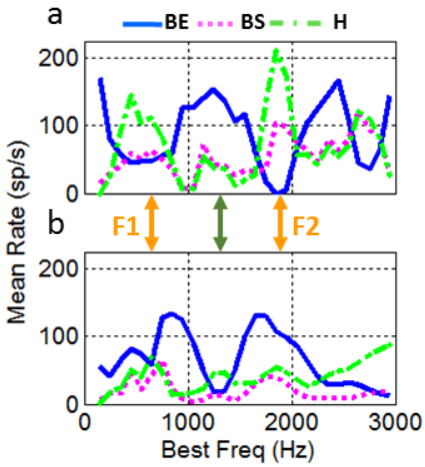


Figure 4. Responses of model IC MTF types to the vowel /æ/ at 65 dB SPL. IC models with (a) healthy AN model inputs, (b) impaired AN inputs. Model parameters for this simulation are provided in Fig. 1.

4 Discussion

The focus of studies of vowel coding has historically been on representation of spectral peaks, or formants, in the rates of AN responses or in the synchrony to stimulus components near BF. These representations are vulnerable to additive noise and vary with sound level. The sensitivity of midbrain neurons to low-frequency fluctuations inspired a shift of focus to the role of the contrast in pitch-related fluctuations along the BF axis in vowel coding (Carney et al., 2015). In the healthy AN, there is a strong and robust contrast in the profile of the F0-synchronized discharge rate across the BF axis (Fig. 2d). Midbrain neurons' sensitivities to these F0-synchronized rates transforms the temporally coded AN profile into a discharge rate profile (Fig. 4a).

The representation of F0-synchronized rates is affected by a number of mechanisms that are influenced by SNHL: synchrony capture, threshold (and thus rate saturation), and bandwidth of tuning. Bandwidth, in turn, affects the modulation spectrum of the peripheral filter outputs; in the case of AN responses, wider peripheral bandwidths associated with SNHL results in a wider modulation bandwidth and a reduction of the amplitude of low-frequency fluctuations in AN responses. These factors don't have a strong influence on the mean rate profile at the level of the AN. However, they have a large effect on AN temporal discharge properties, especially on the F0-synchrony, which in turn affects the responses of IC neurons.

The model responses presented here have implications for the design of signal-processing strategies for listeners with hearing loss. In particular, recreating the spectral representation in

the AN discharge rate profile is not adequate to elicit appropriate responses from central neurons. It is important to re-establish the temporal discharge properties in the AN, in particular the F0-synchronized rate profiles across BF (Rao & Carney, 2014). This result suggests an experimental question: If the profile of peripheral F0-synchrony in response to voiced sounds can be restored to normal for these listeners, can intelligibility of speech, and especially speech in background noise, be improved? The results of such a test would depend on the status of central auditory pathways, which may undergo change following cochlear trauma (Suneja et al., 1998; Salvi et al., 2000).

This modelling study illustrated the substantial effect of a relatively mild hearing loss on the representation of vowels in the midbrain. Future physiological and modelling studies of IC responses to speech should explore the effects of greater hearing loss, as well as environmental noise, reverberation, and their interactions.

Acknowledgments

Supported by NIDCD-010813.

References

Carney, L.H., Li, T., & McDonough, J.M. (2015, in review). Speech Coding in the Brain: Representation of Formants by Midbrain Neurons Tuned to Sound Fluctuations. *eNeuro.*

Delgutte, B., & Kiang, N.Y. (1984). Speech coding in the auditory nerve: I. Vowel-like sounds. *Journal of the Acoustical Society of America,* 75, 866-878.

Deng, L., & Geisler, C.D. (1987). Responses of auditory-nerve fibers to nasal consonant–vowel syllables. *Journal of the Acoustical Society of America,* 82:1977–1988.

Deng, L., Geisler, C.D. & Greenberg, S. (1987). Responses of auditory-nerve fibers to multitone complexes. *Journal of the Acoustical Society of America,* 82:1989–2000.

Fant, G. (1970). *Acoustic Theory of Speech Production.* Hague, The Netherlands: Mouton.

Hillenbrand, J., Getty, L.A., Clark, M.J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America,* 97:3099-3111.

Kewley-Port, D., Burkle, T.Z., & Lee, J.H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America,* 122:2365-2375.

Kim, D.O., Bishop, B.B., Kuwada, S., & Carney, L.H. (2015a). Band-Enhanced and Band-Suppressed Rate Modulation Transfer Functions of Inferior Colliculus Neurons and a Model: Effects of Duty Cycle and Rise/Fall Rate, Association for Research in Otolaryngology. 38th Annual Midwinter Meeting, abstract #PS-423.

Kim, D.O., Zahorik, P., Carney, L.H., Bishop, B.B., & Kuwada, S. (2015b). Auditory Distance Coding in Rabbit Midbrain Neurons and Human Perception: Monaural Amplitude Modulation Depth as a Cue. *The Journal of Neuroscience*, *35*(13), 5360-5372.

Krishna, B.S., & Semple, M.N. (2000). Auditory Temporal Processing: Responses to Sinusoidally Amplitude-Modulated Tones in the Inferior Colliculus. *Journal of Neurophysiology,* 84:255-273.

Langner, G. (1992). A review: Periodicity coding in the auditory system. Hearing Research, 60:115-142.

Miller, R. L., Schilling, J.R., Franck, K.R., & Young, E.D. (1997). Effects of acoustic trauma on the representation of the vowel /e / in cat auditory-nerve fibers. *Journal of the Acoustical Society of America,* 101:3602–3616.

Nelson, P.C., & Carney, L.H. (2004). A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. *Journal of the Acoustical Society of America,* 116:2173-2186.

Rao, A., & Carney, L.H. (2014). Speech Enhancement for Listeners with Hearing Loss Based on a Model for Vowel Coding in the Auditory Midbrain , IEEE Transactions on Bio-medical Engineering. 61:2081-2091.

Sachs, M.B., & Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America,* 66:470-479.

Salvi, R. J., Wang, J., & Ding, D. (2000). Auditory plasticity and hyperactivity following cochlear damage. *Hearing Research*, *147*(1), 261-274.

Studebaker, G.A., Sherbecoe, R.L., McDaniel, D.M., & Gwaltney, C.A. (1999). Monosyllabic word recognition at higher-than-normal speech and noise levels. *Journal of the Acoustical Society of America,* 105:2431-2444.

Suneja, S. K., Potashner, S. J., & Benson, C. G. (1998). Plastic changes in glycine and GABA release and uptake in adult brain stem auditory nuclei after unilateral middle ear ossicle removal and cochlear ablation. *Experimental Neurology*, *151*(2), 273-288.

Young, E.D. (2008). Neural representation of spectral and temporal information in speech. *Philosophical Transactions of the Royal Society of London B: Biological Sciences,* 363:923-945.

Young, E.D., & Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America,* 66:1381-1403.

Zilany, M.S., & Bruce, I.C. (2007). Representation of the vowel/ε/in normal and impaired auditory nerve fibers: model predictions of responses in cats. *The Journal of the Acoustical Society of America*, *122*(1), 402-417.

Zilany, M.S.A., Bruce, I.C., & Carney, L.H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *Journal of the Acoustical Society of America,* 135:283-286.

Chapter 47

Sources of Variability in Consonant Perception and Implications for Speech Perception Modeling

Johannes Zaar and Torsten Dau

Hearing Systems, Department of Electrical Engineering, Technical University of Denmark, Lyngby, Denmark

Contact author: T. D., tdau@elektro.dtu.dk

Abstract

The present study investigated the influence of various sources of response variability in consonant perception. A distinction was made between source-induced variability and receiver-related variability. The former refers to perceptual differences induced by differences in the speech tokens and/or the masking noise tokens; the latter describes perceptual differences caused by within- and across-listener uncertainty. Consonant-vowel combinations (CVs) were presented to normal-hearing listeners in white noise at six different signal-to-noise ratios. The obtained responses were analyzed with respect to the considered sources of variability using a measure of the perceptual distance between responses. The largest effect was found across different CVs. For stimuli of the same phonetic identity, the speech-induced variability across and within talkers and the across-listener variability were substantial and of similar magnitude. Even time-shifts in the waveforms of white masking noise produced a significant effect, which was well above the within-listener variability (the smallest effect). Two auditory-inspired models in combination with a template-matching back end were considered to predict the perceptual data. In particular, an energy-based and a modulation-based approach were compared. The suitability of the two models was evaluated with respect to the source-induced perceptual distance and in terms of consonant recognition rates and consonant confusions. Both models captured the source-induced perceptual distance remarkably well. However, the modulation-based approach showed a better agreement with the data in terms of consonant recognition and confusions. The results indicate that low-frequency modulations up to 16 Hz play a crucial role in consonant perception.

Keywords: talker effects, noise effects, listener effects, internal noise, auditory modeling, perceptual distance, envelope domain, microscopic, perceptual distance, modulation filters.

1 Introduction

Speech perception is often studied from a *macroscopic* perspective, i.e., using meaningful long-term speech stimuli (e.g., in additive noise). To solely investigate the relation between the acoustic properties of the stimulus and the resulting speech percept (excluding lexical, semantic, and syntactic effects), an alternative is to take a *microscopic* perspective by investigating the perception of smaller units of speech such as consonants. Miller & Nicely (1955) measured the perception of consonant-vowel combinations (CVs, e.g., /ba/, /ta/) in white noise and different band-pass filtering conditions and observed distinct consonant confusions. Wang & Bilger

(1973) demonstrated that consonant perception also depends on the vowel context. Phatak & Allen (2007) measured consonant perception in speech-weighted noise and demonstrated noise-type induced perceptual differences to the Miller & Nicely (1955) data. In following studies, perceptual differences across different speech tokens of the same phonetic identity came more into focus (e.g., Phatak et al., 2008).

A few studies have attempted to simulate consonant perception. Li et al. (2010) successfully related consonant recognition data to the so-called AI Gram, which is related to the energy-based Articulation Index (ANSI, 1969). Gallun & Souza (2008) considered noise-vocoded VCVs and demonstrated that the correlation of long-term modulation power representations was a strong predictor of consonant confusions. Jürgens & Brand (2009) used an auditory model in combination with a dynamic time warping (DTW) algorithm that showed convincing recognition predictions while the confusion predictions were inconclusive.

Motivated by the increasing evidence for a major variability in consonant perception that cannot be accounted for by the phonetic identity of the stimuli, the present study attempted to quantify some of the sources of variability that influence consonant perception. It was distinguished between *source-induced variability* and *receiver-related variability*. The former was subdivided into speech- and noise-induced variability; the latter was subdivided into across- and within-listener variability. Consonant perception data were collected and analyzed with respect to the considered sources of variability using a measure of the perceptual distance between responses. Predictions of the data were obtained using an energy- and a modulation-based model in combination with a template-matching back end. The model predictions were compared to the data (i) in terms of how well they reflected the source-induced variability measured in listeners and (ii) in terms of the agreement between perceptual and predicted consonant recognition and confusions.

2 Methods

2.1 Experiment 1: Speech Variability

CVs consisting of the 15 consonants /b, d, f, g, h, j, k, l, m, n, p, s, ʃ, t, v/ followed by the vowel /i/ (as in “feed”) were used. Six recordings of each CV (three spoken by a male, three spoken by a female talker) were taken from the Danish nonsense syllable speech material collected by Christiansen & Henrichsen (2011). Six SNR conditions (12, 6, 0, -6, -12, and -15 dB) were created by fixing the noise level and adjusting the speech level. One particular white masking noise waveform with a duration of one second was generated for each speech token in each SNR condition and mixed with it such that the speech token onset was temporally positioned 400 ms after the noise onset.

2.2 Experiment 2: Noise Variability

Only one male-talker speech token of each CV was used. Three masking-noise conditions (frozen noise A, frozen noise B, and random noise) were considered. For each speech token, one particular white-noise waveform with a duration of one second was generated and labeled ”frozen noise A”; the same noise token was then circularly shifted in time by 100 ms to obtain ”frozen noise B”. The noise tokens were faded in and out using raised cosine ramps with a duration of 50 ms. The noise waveforms for the random noise condition (added to prevent noise learning) were randomly generated during the experimental procedure. The noisy speech tokens were created as described in section 2.1.

2.3 Experimental Procedure

Two different groups of eight normal-hearing native Danish listeners participated in the two experiments (average age: 26 and 24 years, respectively). The stimuli were monaurally

presented to the listeners via headphones in experimental blocks ordered according to the SNR in descending order. An additional quiet condition (clean speech at 60 dB SPL) preceded the SNR conditions (noise level at 60 dB SPL). Each block started with a short training run. The order of presentation within each experimental block was randomized. In experiment 1, each stimulus was presented three times to each listener. In experiment 2, each stimulus was presented five times to each listener. Listeners had to choose one of the response alternatives displayed as 15 buttons labeled "b, d, f, g, h, j, k, l, m, n, p, s, Sj, t, v" and one button labeled "I don't know" on a graphical user interface (the Danish "Sj" corresponds to /ʃ/). Experiment 2 was repeated with four of the originally eight listeners to obtain test-retest data.

2.4 Data Analysis

For each stimulus and listener, the responses obtained in the experiments were converted to proportions of responses by distributing any "I don't know" response evenly across the 15 other response alternatives and dividing the occurrences of responses by the number of stimulus presentations. The response of a given listener obtained with a given stimulus was thus calculated as a vector $r = [p_b, p_d, ..., p_v]$, where p_x denotes the proportion of response "x". The perceptual distance between two response vectors r_1 and r_2 was defined as the normalized angular distance between them:

$$D(r_1, r_2) = \arccos\left(\frac{\langle r_1, r_2 \rangle}{\| r_1 \| \cdot \| r_2 \|}\right) \cdot \frac{100\%}{\pi/2}$$

The perceptual distance was calculated across six different factors: (i) across CVs, (ii) across talkers, (iii) within talkers, (iv) across masking-noise tokens, (v) across listeners, and (vi) within listeners. Apart from the across-CV factor, only responses obtained with stimuli of the same phonetic identity were compared. For each considered factor, the perceptual distance was calculated across all pairwise comparisons of response vectors representative of that factor. The calculations for all factors but (v) were performed for each listener and each SNR condition separately. The calculations for (v) were performed for each SNR condition separately, comparing responses across listeners. The individual distance values were then averaged across the considered response pairs and (where applicable) across listeners. As a result, the respective perceptual distances were obtained as a function of SNR.

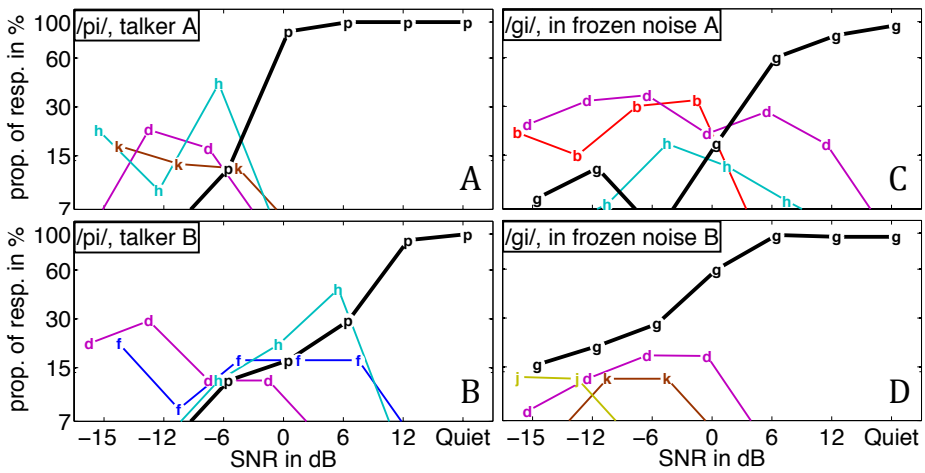


Figure 1. Across-listener average example confusion patterns (CPs). Left: CPs obtained with two different speech tokens /pi/ spoken by male talker A (top) and female talker B (bottom). Right: CPs obtained with one specific speech token /gi/ mixed with frozen noise A (top) and frozen noise B (bottom).

3 Experimental Results

Figure 1 shows examples of perceptual *across-talker variability* and perceptual *across-noise variability* in terms of across-listener average confusion patterns: /pi/ spoken by talker A (panel A) was more recognizable (and less confusable) than /pi/ spoken by talker B (panel B); the perception of a given speech token /gi/ was very differently affected by a specific white masking-noise waveform "A" (panel C) than by a time-shifted version of that waveform ("B", panel D).

Figure 2 shows the perceptual distances derived from the experimental data for all SNR conditions (see Section 2.4). On the left, the across-SNR average is shown. The largest perceptual distance of 91% was found across CVs (black bar). Regarding the source-induced perceptual distances across stimuli of the same phonetic identity, the largest perceptual distance of 51% was obtained across talkers (blue bar), followed by the within-talker factor (47%, green bar). A temporal shift in the masking-noise waveform induced a perceptual distance of 39% (red bar). Regarding the receiver-related effects, a substantial perceptual distance of 46% across listeners was found for physically identical stimuli (light grey bar). In contrast, the relatively low perceptual distance of 30% within listeners (dark grey bar) indicated that the individual listeners were able to reproduce their responses fairly reliably. Pairwise t-tests across all combinations of conditions (excluding the across-CV condition) demonstrated that all conditions were significantly different from each other ($p < 0.05$) except for the across-talker (blue), within-talker (green), and across-listener (light grey) conditions.

Regarding the trends across SNR in Figure 2, it can be seen that the across-CV distance (black bars) was at ceiling for large SNRs and decreased with decreasing SNR, as listeners made more speech-token specific confusions. All other perceptual distance types showed low values for large SNRs and increased with decreasing SNR due to stimulus-specific confusions and listener uncertainty. The within-listener distance ("internal noise") represented the baseline and strongly increased with decreasing SNR as the task became more challenging.

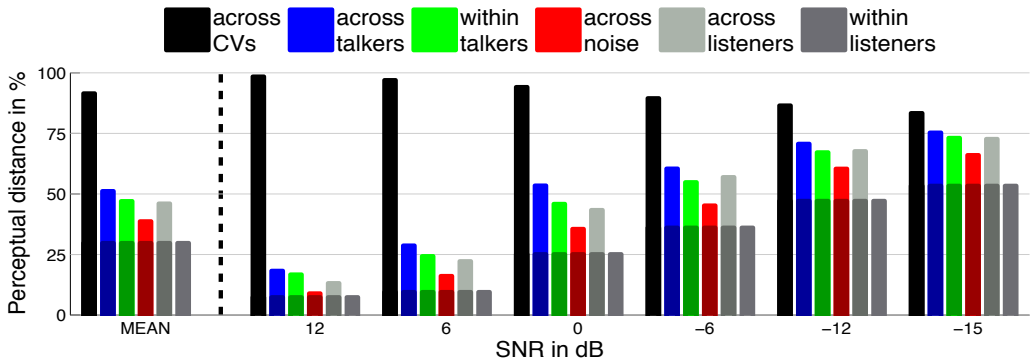


Figure 2. Mean (left) and SNR-specific perceptual distances across CVs, across talkers, within talkers, across noise, across listeners, and within listeners. The shaded areas represent values below the within-listener distance, i.e., below the internal-noise baseline.

4 Modeling

4.1 Model Components

The subband power P, in dB, was calculated using 22 gammatone filters with equivalent rectangular bandwidths. The center frequencies were spaced on a third-octave grid, covering a range from 63 Hz to 8 kHz. The Hilbert envelope of each filter output was extracted and low-pass filtered using a first-order Butterworth filter with a cut-off frequency of 150 Hz. The envelopes were down-sampled to a sampling rate of 1050 Hz.

The modulation power P_{mod}, in dB, was obtained using the subband envelope extraction described above, followed by a modulation filterbank consisting of three second-order band-pass filters (center frequencies of 4, 8, and 16 Hz) in parallel with one third-order low-pass filter (cut-off frequency of 2 Hz).

A template-matching procedure was applied to predict the responses obtained in experiment 1. Two talker-specific template sets were considered, consisting of all speech tokens from each talker (i.e., three templates for each CV). The templates were mixed with random white noise at the test-signal SNR and compared to the experimental stimuli. The distances between the models' internal representations of the test signals and the templates were obtained using a standard dynamic time warping (DTW) algorithm (Sakoe & Chiba, 1978). The template-matching procedure was conducted nine times with newly generated random noise for the templates. The "correct" template always contained the same speech token as the test signal, while the masking noise differed. In each run, the template showing the smallest distance to the test signal was selected. The modeled responses were converted to proportions of responses. The responses obtained in experiment 2 were predicted similarly, considering only the 15 speech tokens used in experiment 2 as templates.

4.2 Simulation Results

Sources of Variability

In accordance with the procedure described in Section 2.4, the across-CV, across-talker, within-talker, and across-noise modeled distances were obtained as a function of the SNR from the predicted responses. Figure 3 shows scatter plots of the perceptual distance versus the modeled distances obtained using P (panel A) and P_{mod} (panel B). It can be observed that the perceptual distances were remarkably well-predicted using P as well as P_{mod}, with a Pearson's *r* of 0.96 and 0.98, respectively.

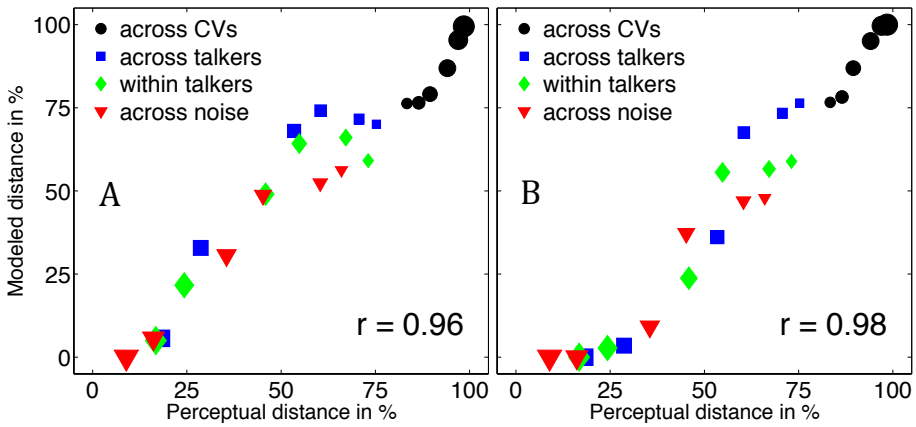


Figure 3. Source-induced perceptual distance shown in Figure 2 plotted versus simulation results obtained using P (left) and P_{mod} (right). The symbols and colors represent the different distance types, the size of the symbols is proportional to the SNR.

Consonant Recognition and Confusions

The grand average consonant recognition rates obtained in experiment 1 showed a speech reception threshold (SRT) of -3 dB. The predicted SRTs were overestimated by 2.8 dB using P and by only 0.4 dB using P_{mod}. The token-specific SRTs showed a large spread across speech tokens, which was smaller in both model predictions. However, P_{mod} captured the relative ranking of token-specific SRTs considerably better than P (Spearman's *r* of 0.4 and 0.04, respectively).

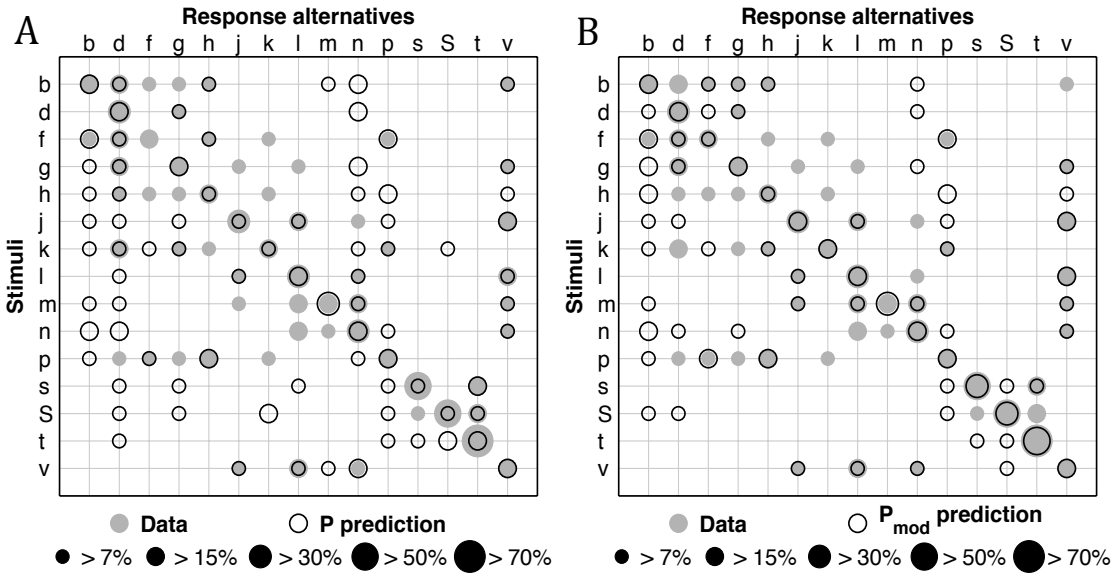


Figure 4. Confusion matrices obtained in experiment 1, averaged across listeners, speech tokens of the same phonetic identity, and across SNRs of 0, -6, -12, and -15 dB. The perceptual data are shown as filled grey circles in both panels. The model predictions obtained with P (panel A) and P_{mod} (panel B) are represented as open black circles.

The across-listener average data obtained in experiment 1 were averaged across different speech tokens of the same phonetic identity and across the four lowest SNRs (as most confusions occur for low SNRs). The resulting confusion matrices (CMs) are shown as filled grey circles in both panels of Figure 4. The model predictions are plotted on top as open black circles. For P (panel A), an underestimation of the recognition for many consonants was observed, indicated by the mismatch of the on-diagonal circles. For P_{mod} (panel B), a good consonant-specific recognition match was found. Both models hit most of the confusions, reflected in the proportion of grey off-diagonal circles matched with a black circle. However, there were also many "false alarms", particularly in the P-based predictions.

5 Summary and Discussion

The investigation of different sources of variability in Section 3 indicated that any acoustical difference in the stimuli produced a measurable effect. The observed perceptual variability across talkers is well established in the related literature (e.g., Phatak et al., 2008); the equally large variability within talkers had not yet been demonstrated. Most remarkably, even a 100-ms time shift in the white masking-noise waveform induced significant perceptual variability, indicating that "steady-state" masking noise should not be considered steady over time in the context of consonant cues. On the receiver side, different NH listeners with identical language background showed substantial differences while individual listeners could fairly reliably reproduce their responses. Averaging consonant perception data (even across NH listeners) thus seems problematic.

The predictions obtained in Section 4 with the energy-based (P) and the modulation-based (P_{mod}) pre-processing stages both accounted for the trends in the perceptual data with respect to the considered stimulus-related sources of variability. Consonant recognition was strongly underestimated using P and well-explained using P_{mod}. An inspection of confusion matrices suggested that both models correctly predicted most of the perceptual confusions, albeit with some "false alarms". The overall larger predictive power obtained with P_{mod} indicates that slow envelope fluctuations up to 16 Hz are a good predictor for consonant-in-noise perception. This is consistent with the findings by Gallun & Souza (2008).

The perceptual data analysis has implications for the further model design. It was shown that the internal noise increased with decreasing SNR. This could be incorporated in the model using an SNR-dependent random process in the decision stage. Furthermore, the model predicted responses of a hypothetical "average" NH listener, which is unrealistic given the considerable across-listener variability. It remains a challenge to include listener-specific differences in the model, as it is not clear whether these differences can be accounted for by slight audiometric differences between the NH listeners, cognitive effects, individual biases, or any combination of these factors. Eventually, an extension of the model towards different types of hearing impairment might be useful to understand the link between individual impairment factors and microscopic speech intelligibility.

Acknowledgments

We thank Søren Jørgensen for his contributions to the research underlying this study. This research was supported by the FP7 Marie Curie Initial Training Network INSPIRE.

References

ANSI (1969). ANSI S3.5-1969 American national standard methods for the calculation of the articulation index. Standards Secretariat, Acoustical Society of America.

Christiansen, T. U., & Henrichsen, P. J. (2011). Objective evaluation of consonant-vowel pairs produced by native speakers of Danish. *Proceedings of Forum Acusticum 2011*. http://openarchive.cbs.dk/bitstream/handle/10398/8412/Peter_Juel_Henrichsen_2.pdf?sequence=1. Accessed 16 April 2015.

Gallun, F., & Souza, P. (2008). Exploring the role of the modulation spectrum in phoneme recognition. *Ear and Hearing*, 29(5), 800–813.

Jürgens, T., & Brand, T. (2009). Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America*, 126(5), 2635–2648.

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, 127(4), 2599–2610.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27(2), 338–352.

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, 121(4), 2312–2326.

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, 124(2), 1220–1233.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Proc.* ASSP-26(1), 43–49.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *The Journal of the Acoustical Society of America*, 54(5), 1248–1266.

Chapter 48

On Detectable and Meaningful Speech-Intelligibility Benefits

William M. Whitmer, David McShefferty, Michael A. Akeroyd
MRC/CSO Institute of Hearing Research – Scottish Section, Glasgow, UK

Contact author: WMW, bill@ihr.gla.ac.uk

Abstract

The most important parameter that affects the ability to hear and understand speech in the presence of background noise is the signal-to-noise ratio (SNR). Despite decades of research in speech intelligibility, it is not currently known how much improvement in SNR is needed to provide a meaningful benefit to someone. We propose that the underlying psychophysical basis to a meaningful benefit should be the just noticeable difference (JND) for SNR. The SNR JND was measured in a series of experiments using both adaptive and fixed-level procedures across participants of varying hearing ability. The results showed an average SNR JND of approximately 3 dB for sentences in same-spectrum noise. The role of the stimulus and link to intelligibility was examined by measuring speech-intelligibility psychometric functions and comparing the intelligibility JND estimated from those functions with measured SNR JNDs. Several experiments were then conducted to establish a just meaningful difference (JMD) for SNR. SNR changes that could induce intervention-seeking behaviour for an individual were measured with subjective scaling and report, using the same stimuli as the SNR JND experiment as pre- and post-benefit examples. The results across different rating and willingness-to-change tasks showed that the mean ratings increased near linearly with a change in SNR, but a change of at least 6 dB was necessary to reliably motivate participants to seek intervention. The magnitude of the JNDs and JMDs for speech-intelligibility benefits measured here suggest a gap between what is achievable and what is meaningful.

Keywords: signal-to-noise ratio, just-noticeable difference, speech intelligibility, hearing impairment

1 Introduction

The inability to provide appreciable speech-in-noise benefits can lead to the non-use of hearing aids (e.g., McCormack & Fortnum, 2013). Non-use can come from unmet expectations (i.e., benefits without satisfaction; Demorest 1984), which manifest as meaningless benefits or simply undetectable benefits. The current study specifically looks at discriminable and meaningful speech intelligibility benefits based on changes in signal-to-noise ratio (SNR) when the signal is speech. Previously there has been only one small study related to measuring the just-noticeable difference (JND) in SNR by Killion (2004), in which hearing-aid wearers were at chance discriminating a 2 dB SNR difference for speech in noise. The description of that study, however, was limited; a more rigorous testing of the SNR JND was performed in the current study across a broad range of hearing abilities and ages.

A JND, though, is just a psychophysical benchmark. How large does a change in speech in noise need to be for it to be meaningful to someone? What is meaningful and clinically significant is often applied to service-wide treatments, and could be vital for determining provision criteria, as well as tempering expectations for the patient. The current study looks at what could induce intervention-seeking behaviour for an individual; that is, how much subjective value do patients ascribe to discriminable speech intelligibility benefits? Using the same stimuli as examples of pre- and post-benefit situations, the current study also measured what is a meaningful benefit: the minimum SNR change necessary to spur someone to seek out a clinical intervention (e.g., a new or adjusted hearing aid).

2 Detectable Benefits

Based on the premise that a meaningful benefit should be a reliably discriminable benefit, we measured using multiple procedures as described below, what is a discriminable benefit: the just noticeable difference in SNR.

2.1 Changes in SNR

As part of several experiments, the JND for SNR was measured for adults of varying hearing ability at different reference SNRs using adaptive and fixed-level procedures based on the classic level discrimination paradigm (cf. McShefferty et al., 2015). Across these experiments, the stimuli were equalised IEEE sentences uttered by a native British English speaker (Smith & Faulkner, 2006) in same-spectrum noise, presented over headphones in a sound-attenuated booth. Noises began and ended simultaneous with the speech signal, and there was a 500-ms gap between intervals. Both intervals contained the same randomly chosen sentence, and after both intervals, participants were prompted to respond to which one was clearer. In each interval the levels of both speech and noise were adjusted to maintain a presentation level of 75 dB A; to minimize other cues, the level of each interval was roved independently by a maximum of ±2 dB. In a two-interval/alternative forced-choice task, participants were presented, in randomized order, with a reference interval at a fixed reference SNR, and a target interval at the reference SNR plus an increment (ΔSNR) that was either varied adaptively or presented at fixed, randomly interleaved ΔSNRs. The adaptive procedure estimated 79% correct from the geometric means of the best two of three three-up/one-down tracks. The fixed-level procedure estimated 79% correct from fitting a maximum-log-likelihood logistic function to the data. The procedures produced equivalent results to 0.1 dB, so results were combined across procedures for each reference SNR.

Collated results are shown in Table 1. There were three SNR references in all: -6, 0 and +6 dB SNR. The 2.8 and 2.9 dB SNR JNDs for -6 and 0 dB reference SNRs were equivalent, and both SNR JNDs were significantly less than the 3.7 dB SNR JND for the +6 dB SNR reference condition [$t_{(91.3)} = 4.07$; $p = 0.0001$ and $t_{(121.4)} = 4.11$; $p = 0.00007$, respectively].

Table 1. Mean (μ) SNR JND for reference SNRs of -6, 0 and +6 dB, showing standard deviation (σ) and participant number (*n*) tested at each reference SNR. Median better-ear four-frequency pure-tone averages (BE4FAs) and ages are given with ranges in parentheses.

	-6 dB	0 dB	+6 dB
μ (dB SNR)	2.8	2.9	3.7
σ	1.0	1.0	1.4
n	35	99	72
BE4FA (dB HL)	28 (3-56)	29 (-1-85)	32 (-1-71)
Age (yrs)	63 (22-72)	65 (23-76)	64 (22-76)

2.2 Changes in Intelligibility

The JND for intelligibility was estimated by first measuring percentage correct for IEEE sentences based on keywords at SNRs of -16-0 or -8-8 dB in 4 dB increments. Twenty-four adult participants, median better-ear four-frequency pure-tone average (BE4FA) of 16 dB HL (range -3-49 dB HL) and median age of 57 (range 27-74), were first presented with sentences at -16 and 0 dB SNR. 14 participants could not respond with half of the keywords (5 keywords/sentence) at 0 dB; those participants were tested at the -8-8 dB SNR range. The individual results, averaged across 50 sentences (250 keywords) at each SNR, were fit with a maximum-log-likelihood logistic function. The median SNR corresponding to 50% correct on each individual's psychometric function (SNR50) was 0.20 dB and ranged from -3.9 to 5.3 dB. The SNR JND was then measured adaptively with SNR50 as the reference. The mean SNR JND was 3.0 dB (σ = 1.0 dB), equivalent to the SNR JNDs for -6 and 0 dB SNR, and also significantly different from the SNR JND for +6 dB SNR [$t_{(57.8)} = 2.86$; $p = 0.006$]. The reference SNR + SNR JND value (in dB) was then converted to 50% + Intelligibility JND (in %) using the slope of the fit to the speech intelligibility data. The mean Intelligibility JND estimate for IEEE sentences was 26.4% (σ = 7.5%).

To investigate whether the JNDs initially measured (Section 2.1 above) were possibly Intelligibility rather than SNR JNDs, as well as to probe the effect of stimulus, the same procedure was repeated for digit triplets presented in white noise. The hypothesis was that, if it was indeed an SNR JND, the JND differences across stimuli should be greater for Intelligibility than SNR. The mean SNR JND for digits was 2.5 dB (σ = 0.8 dB); the mean Intelligibility JND estimate for digits was 17.5% (σ = 7.2%). SNR and Intelligibility JNDs were significantly lower for digits in unfiltered random (white) noise than for sentences in same-spectrum noise [$t_{(23)} = 2.98$; $p = 0.007$ and $t_{(23)} = 5.00$; $p = 0.00005$]. Across stimuli, both the SNR JND differences (mean 0.5 dB) and the Intelligibility JND differences (mean 8.92 dB) were significantly non-zero [$z_{(23)} = 2.98$; $p = 0.003$, and $z_{(23)} = 5.00$; $p \approx 0$]. Hence, it is not certain whether the 3 dB SNR JND is indeed the JND for SNR or for Intelligibility.

3 Meaningful Benefits

To ascertain what is a meaningful benefit to someone requires more than detectability; it requires the subjective input of the participant. The previous method for deriving the SNR JND was hence changed to subjective-comparison tasks using SNR and SNR + ΔSNR pairs as examples of situations to measure the just meaningful difference (JMD) for SNR. It is not clear what query, however, best represents meaningfulness. Four tasks are discussed below: a conventional better/worse rating task, a novel conversation-tolerance rating task, a swap-paradigm task and a clinical significance task. For the latter two tasks (Sections 3.3 and 3.4), the JMD was considered as the ΔSNR where the proportion of affirmative responses (i.e., willingness to swap devices or attend the clinic) were significantly greater than chance.

3.1 Rateable Benefits

The simplest, most direct rating of preference is a better/worse scaling procedure. Thirty six participants (median age: 62 years; range: 31-74) of varying hearing ability (median BE4FA: 21 dB HL; range 3-85 dB HL) heard two intervals of IEEE sentences in same-spectrum noise, one at 0 dB SNR, and the other at 0 dB plus a ΔSNR of 1, 2, 4, 6 or 8 dB. The overall level of each interval was 75 dB A; unlike the JND experiments, the level was not roved across intervals. Participants were asked how much better/worse the second interval was on a discrete, signed 11-point scale with -5 marked "much worse" and +5 marked "much better." Each of the ten conditions (5 ΔSNRs × 2 orders) was repeated 12 times in randomized order, resulting in 120 trials completed in three blocks of 40.

Results, including 95% within-subject confidence intervals, are shown in the left panel of Figure 1. Ratings increase near linearly as a function of ΔSNR, and are asymmetric: when the +ΔSNR interval preceded the reference, ratings were less than when the +ΔSNR interval followed the reference (e.g., when the second interval was the better SNR, ratings were greater than one increment at a ΔSNR of 4 dB, but when the second interval was the worse SNR, ratings were greater than one increment only when ΔSNR was 8 dB. This asymmetry was most likely due to an order effect; the second iteration of a sentence has been shown to be more intelligible (e.g., Thwing, 1956), hence second-interval deficits would be deflated, and second-interval benefits inflated. Regarding the JMD, the ratings significantly increased only at 4 dB, regardless of order, lending support to the notion of detectability being a requisite for (arbitrary) meaning. SNR JNDs for these participants, however, were not correlated with their individual mean responses at any ΔSNR.

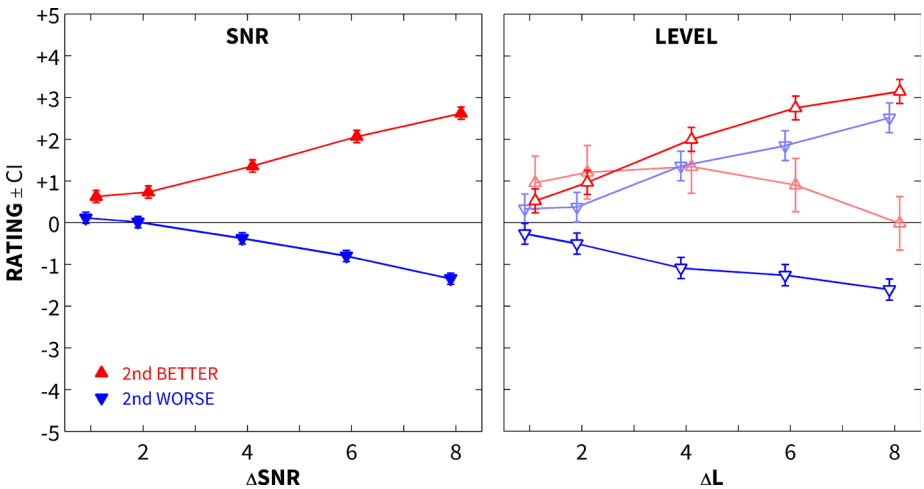


Figure 1. Better (+) and worse (-) ratings as function of changes in SNR (left panel) and overall level (right panel). The reference SNR was 0 dB; the reference level was 70 dB A. Faded symbols/lines (right panel) indicate mean ratings for six participants whose max./min. ratings did not occur at the extremities. Error bars show 95% within-subject confidence intervals (based on the analysis of variance subject × condition interaction term).

An analogous procedure was used to examine what ratings would apply to a global change in level as opposed to a change in SNR. Thirty-six participants (median age: 67 years; range: 27-77) of varying hearing ability (median BE4FA: 37 dB HL; range: 4-84) heard stimuli at a fixed SNR of 0 dB. The reference level (L) was 70 dB A, and ΔL on any trial was 1, 2, 4, 6 or 8 dB. The level JND was also measured for each participant; the mean was 1.4 dB (σ = 0.4 dB), similar to previous level JNDs for older hearing-impaired adults (Whitmer & Akeroyd, 2011).

Results are shown in the right panel of Figure 1. Six participants exhibited varying negative reactions to the greatest levels, finding a particular non-maximal level rated highest (e.g., 66 and 74 dB A). For the remaining 30 participants, the results are very similar to those for SNR, though

ΔLs of 4-8 dB were rated modestly higher than ΔSNRs of 4-8 dB. That is, a change in level without a change in SNR was considered as good as – if not better than a change in SNR. While this finding does not coincide with their role in intelligibility, it is coincident with detectability of level and SNR changes (JNDs of 1.5 and 3.0 dB, respectively).

3.2 Tolerable Benefits

A novel method for establishing a meaningful difference was developed based around complaints of not being able to endure noisy conversation. For the conversation tolerance test, listeners heard the same sentence at two SNRs, one at an estimate of their speech-in-noise reception threshold (SNR50) and one at SNR50 plus an increment of 0.5-8 dB. In each of a pair of intervals, the participant was asked, "How many conversations in this situation would you tolerate?" (i.e., a paired single-interval task). Twenty-one adults (median age: 66 years; range: 52-73) with varying hearing ability (median BE4FA: 24 dB HL; range: 4-61) completed the task, repeating each increment in randomized order ten times.

Results are shown in Figure 2. On average, participants did not tolerate one additional conversation until the SNR increased by at least 3 dB, and was not significantly greater than one conversation (a less arbitrary unit than the better/worse rating) until a 4 dB SNR change, similar to the results in Section 3.1. Despite responding after each presentation, having the same sentence presented twice created an order effect: when the reference SNR was presented after the greater SNR for the same sentence, the tolerance rating increased on average 1.8 conversations. Unlike the ratings experiment, though, the increase in conversation toleration is clearly linear as a function of SNR change, not indicating any relationship to the SNR JND.

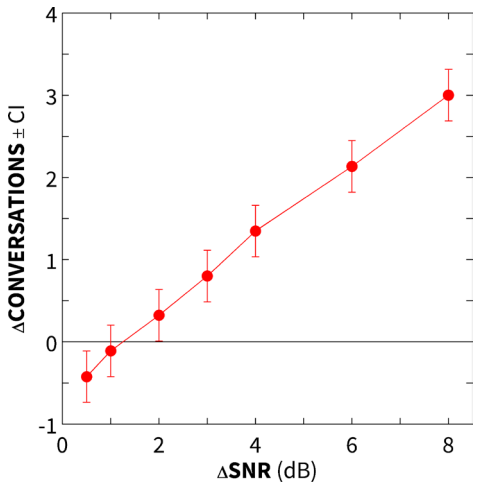


Figure 2. Change in conversations tolerated as a function of the change in SNR. Error bars show 95% within-subject confidence intervals.

3.3 Swappable Benefits

If another hearing aid offers more speech-in-noise benefit, would someone be willing to trade for it? Using the same stimuli as the rating experiment (3.1), 35 participants (median age: 62 years; range: 38-74) of varying hearing ability (median BE4FA: 34 dB HL; range: 4-80) were asked to consider the reference SNR (either -6 or +6 dB) interval as an example of their current device, and the reference SNR + ΔSNR (2, 4, 6 or 8 dB) interval as a different device, and were asked on each trial if they would swap their current for a different device to get that change (yes/no). Each condition (reference SNR × ΔSNR) was repeated 30 times in randomized order for a total of 240 trials.

Results are shown in the left panel of Figure 3. When the initial SNR was -6 dB, the proportion of times that participants, on average, were willing to swap exceeded chance only when the ΔSNR

was greater than 4 dB (i.e., 70% at 6 dB SNR, and 87% at 8 dB SNR). When the initial SNR was +6 dB, willingness to swap just exceeded chance (56% willing to swap) at a ΔSNR of 8 dB. The JMD based on the swap paradigm is therefore dependent on the difficulty of the situation (i.e., the reference SNR), but appears to have a minimum of 6 dB SNR.

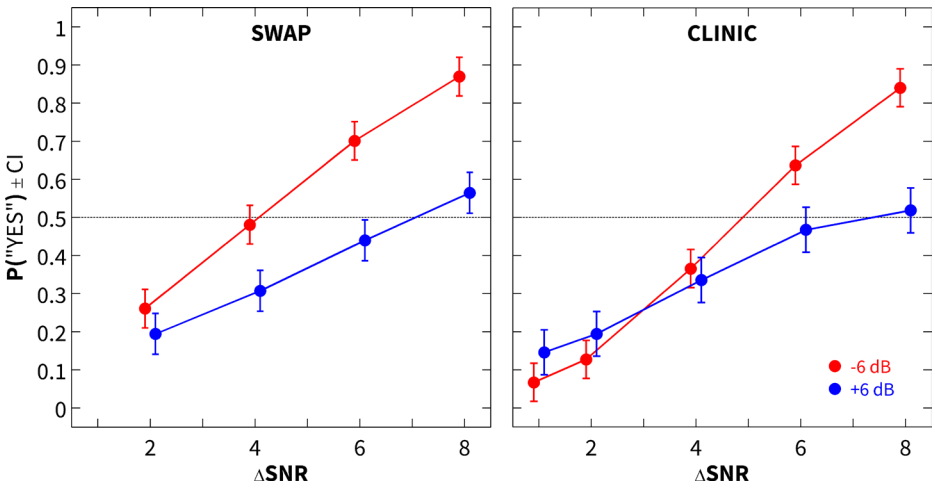


Figure 3. Proportion of yes responses for willingness to swap devices (left panel) and willingness to attend the clinic (right panel) as a function of positive change in SNR (ΔSNR) for two reference SNRs: -6 (filled circles) and +6 dB SNR (open circles). Error bars show 95% within-subject confidence intervals

3.4 Clinically Significant Benefits

While there are statistical bases for a minimal clinically important difference (cf. Jaeschke et al., 1989), the goal here was to develop a benchmark for what speech intelligibility benefit is necessary to motivate an individual to seek intervention – a just meaningful difference (JMD). Hence, clinical significance was applied in a literal sense: participants were asked whether a positive change in SNR, as an example of what a visit to the clinic would provide, was worth attending the clinic to get that change. Thirty-six participants (median age: 63 years; range: 22-72) of varying hearing ability (median BE4FA: 28 dB HL; range: 2-56) were presented stimuli as in Section 3.1 and 3.3 with 12 trials of each reference SNR (-6 and +6 dB) and each ΔSNR (1, 2, 4, 6 and 8 dB) for a total of 120 trials.

Results are shown in the right panel of Figure 3. The results were similar to the swap experiment (left panel of same figure), though there was a greater tendency towards not attending the clinic compared to not swapping devices at the lowest, sub-JND ΔSNRs. For the -6 dB SNR reference condition, the mean proportion of yes responses only exceeded chance at 6 and 8 dB SNR. For the +6 dB SNR reference condition, mean responses never significantly exceeded chance; furthermore, the function appears asymptotic, so may not have exceeded chance regardless of ΔSNR. If the situation is already advantageous, making it clearer does not appear to induce intervention-seeking behaviour. The JMD derived from clinical significance is the same as the JMD derived from the swap task: a minimum of 6 dB SNR.

4 Conclusions

To determine both detectable and meaningful speech-intelligibility benefits, participants were presented over several experiments paired examples of speech and noise, one at a reference SNR and the other at a variably higher SNR. The threshold for a discriminable benefit was roughly 3 dB SNR. In more advantageous conditions, the SNR JND increased, and decreased for simpler speech (digits vs. sentences; cf. Macpherson & Akeroyd, 2014). The threshold for a meaningful benefit was at least 6 dB. Curiously, age and hearing loss were not factors well

correlated with either individuals' just noticeable or just meaningful differences. At least 6 dB – more in less noisy situations – is necessary to motivate participants to seek an intervention based solely on changes in speech and noise levels. Given the difficulty achieving SNR benefits greater than 6 dB in realistic environments with current hearing aid processing outwith wireless transmission of the signal, the evidence here indicates that there is no demonstrable "wow" in speech-intelligibility benefits. Rather, assessing other potential avenues of benefit – such as long-term improvements, attentional/cognitive ease, and psycho-social engagement – could show the impact of sub-JND/JMD speech intelligibility benefits.

Acknowledgments

We would like to thank Dr. Michael Stone for supplying some of the stimuli used. The Scottish Section of IHR is supported by intramural funding from the Medical Research Council (grant number U135097131) and the Chief Scientist Office of the Scottish Government.

References

Demorest, M. E. (1984). Techniques for measuring hearing aid benefit through self-report. In J. Pickett, ed., Symposium on hearing aid technology: Its present and future, Gallaudet, Washington, DC.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. Controlled Clinical Trials, 10, 407–415.

Killion, M. C. (2004). Myths about hearing in noise and directional microphones. Hearing Review, 11, 14-19, 72-73.

Macpherson, A., & Akeroyd, M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. Trends in Hearing, 18, 1–26.

McCormack, A., & Fortnum, H. (2013). Why do people with hearing aids not wear them? International Journal of Audiology, 52, 360-368.

McShefferty, D., Whitmer, W. M., & Akeroyd, M. A. (2015). The just-noticeable difference in speech-to-noise ratio. Trends in Hearing, 19, 1-9. doi:10.1177/2331216515572316

Smith, M. W., & Faulkner, A. (2006). Perceptual adaptation by normally hearing listeners to a simulated 'hole' in hearing. Journal of the Acoustical Society of America, 120, 4019–4030.

Thwing, E. J. (1956). Effect of repetition on articulation scores for PB words. Journal of the Acoustical Society of America, 28, 302–303.

Whitmer, W. M., & Akeroyd, M. A. (2011). Level discrimination of speech sounds by hearing-impaired individuals with and without hearing amplification. Ear and Hearing, 32, 391–398.

Chapter 49

Individual Differences in Behavioral Decision Weights Related to Irregularities in Cochlear Mechanics

Jungmee Lee, Inseok Heo, An-Chieh Chang, Kristen Bond, Christophe Stoelinga, Robert Lutfi
Communication Sciences and Disorders, University of Wisconsin-Madison, Madison, WI, USA

Glenis R. Long
Speech and Hearing Sciences, Graduate Center of City University of New York, New York, NY, USA

Contact author: GL, glong@gc.cuny.edu

Abstract

An unexpected finding of previous psychophysical studies is that listeners show highly replicable, individualistic patterns of decision weights on frequencies affecting their performance in spectral discrimination tasks – what has been referred to as *individual listening styles* (Doherty and Lutfi, 1996; Lutfi and Liu, 2007; Jesteadt et al., 2014). We, like many other researchers, have attributed these listening styles to peculiarities in how listeners attend to sounds, but we now believe they partially reflect *irregularities* in cochlear micromechanics affecting what listeners hear. The most striking evidence for cochlear irregularities is the presence of low-level spontaneous otoacoustic emissions (SOAEs) measured in the ear canal and the systematic variation in stimulus frequency otoacoustic emissions (SFOAEs), both of which result from back-propagation of waves in the cochlea. SOAEs and SFOAEs vary greatly across individual ears and have been shown to affect behavioural thresholds, behavioural frequency selectivity and judged loudness for tones (Long and Tubis, 1988, Baiduc, Lee and Dhar, 2012, Mauermann, Long and Kollmeier, 2004). The present paper reports pilot data providing evidence that SOAEs and SFOAEs are also predictive of the relative decision weight listeners give to a pair of tones in a level discrimination task. In one condition the frequency of one tone was selected to be near that of an SOAE and the frequency of the other was selected to be in a frequency region for which there was no detectable SOAE. In a second condition the frequency of one tone was selected to correspond to an SFOAE maxima, the frequency of the other tone, an SFOAE minima. In both conditions a statistically significant correlation was found between the average relative decision weight on the two tones and the difference in OAE level.

Keywords: behavioural decision weights, level discrimination, spontaneous otoacoustic emissions, stimulus frequency otoacoustic emission.

1 Introduction

People with normal hearing acuity usually can follow a conversation with their friends at a noisy party, a phenomenon known as the “cocktail party effect” (Cherry, 1953). This remarkable ability to attend to target sounds in background noise deteriorates with age and hearing loss. Yet, people who have been diagnosed in the clinic as having a very mild hearing loss or even

normal hearing based on their pure tone audiogram (the clinical gold standard for identifying hearing loss) still often report considerable difficulty communicating with others in such noisy environments (King and Stephens, 1992). The conventional pure tone audiogram, the cornerstone of hearing loss diagnosis, is not always the best predictor for these kinds of difficulties.

Perturbation analysis has become a popular approach in psychoacoustic research to measure how listeners hear out a target sound in background noise (cf. Berg, 1990; Lutfi,, 1995; Richards, 2002). Studies using this paradigm show listeners to have *highly replicable, individualistic patterns* of decision weights on frequencies affecting their ability to hear out specific targets in noise – what has been referred to as *individual listening styles* (Doherty & Lutfi 1996; Lutfi and Liu, 2007; Jesteadt et al., 2014; Alexander and Lutfi, 2004). Unfortunately this paradigm is extremely time-consuming, rendering it ineffective for clinical use. Finding a quick and an objective way to measure effective listening in noisy environments would provide a dramatic improvement in clinical assessments, potentially resulting in better diagnosis and treatment.

In the clinic, otoacoustic emissions (OAEs) provide a fast, noninvasive means to assess auditory function. Otoacoustic emissions (OAEs) are faint sounds that travel from the cochlea back through the middle ear and are measured in the external auditory canal. Since their discovery in the late 1970s by David Kemp (1978), they have been used in the clinic to evaluate the health of outer hair cells (OHC) and in research to gain scientific insight into cochlear mechanics. Behaviorally they have been shown to predict the pattern of pure-tone quiet thresholds (Long & Tubis, 1988; Dewey & Dhar, 2014), auditory frequency selectivity, (Baiduc et al. 2014), and loudness perception (Mauermann et al., 2004).

The effect of threshold microstructure (and thus OAEs) on loudness perception is particularly noteworthy because relative loudness is also known to be one of the most important factors affecting the decision weights listeners place on different information-bearing components of sounds (Berg, 1990; Lutfi & Jesteadt, 2006; Epstein & Silva, 2009; Thorson 2012; Rasetshwane et al., 2013). This suggests that OAEs might be used to diagnose difficulty in target-in-noise listening tasks through their impact on decision weights. OAEs may be evoked by external sound stimulation (EOAEs) or may occur spontaneously (SOAEs). Stimulus frequency OAEs (SFOAEs), which are evoked using a single frequency sound, are one of the most diagnostic OAEs regarding cochlear function. They show a *highly replicable individualistic pattern* of amplitude maxima and minima when measured with high enough frequency resolution, a pattern called SFOAE fine structure. The level difference between maxima and minima can be as large as 30 dB. Usually SOAEs occur near the maxima of SFOAEs fine structure (Bergevin et al., 2012; Dewey & Dhar, 2014). Given that loudness varies with SFOAE maxima and minima and that loudness is a strong predictor of listener decision weights, it is possible that both SFOAEs and SOAEs may be used to predict individual differences in decision weights.

2 Methods

2.1 Listeners

Data is presented from seven individuals (mean age: 27.42 yrs old) with pure tone air-conduction hearing thresholds better than 15 dB HL at all frequencies between 0.5 and 4 kHz, normal tympanograms, and no history of middle ear disease or surgery.

2.2 Measurement and Analysis of Otoacoustic Emissions

SOAEs were evaluated from 3 minute recordings of sound in the ear canal obtained after subjects were seated comfortably for 15 minutes in a double-walled, Industrial Acoustics, sound-attenuation chamber. The signal from the ER10B+ microphone was amplified by an Etymotic preamplifier with 20dB gain before being digitized by Fireface UC (16bit, 44100 samples/sec).

The signal was then segmented into 1-sec analysis windows (1-Hz frequency resolution) with a step size of 250ms. Half of the segments with the highest power were discarded in order to reduce the impact of subject generated noise. Then an estimate the spectrum in the ear canal was obtained by converting the averages of FFT magnitude in each frequency bin to dB SPL. SOAE frequencies were determined when there was a very narrow peak of energy at least 3 dB above an average of the background level of adjacent frequencies.

SFOAE measurements were obtained using an adaptation of a swept-tone paradigm (cf. Long & Talmadge, 2007, 2008). A probe tone was swept logarithmically at a rate of 2 seconds per octave from 500Hz to 8000Hz at 35dB SPL (8 sec per sweep). The stimuli were delivered to the subjects' ear canal through ER2 tube phones attached to an ER10B+ probe microphone used to record the OAEs. The sweep was repeated 32 times to permit reduction of the noise and intermittent artifacts. Each measurement was multiplied by a half-second-long sliding window of a step size of 25ms, which gives successive windowed signal segments x_m^n where n = 1,2,...,32 and m are the indices of repeated number and segments respectively. An artifact-reduced segment $\bar{x}_m$ was obtained by taking a temporal average of the 32 segments having power falling within 75% of the median, then the magnitude coefficients $\bar{X}_m$(at frequency $F_m = 500 \cdot 2^{(0.025/2)m}$) were evaluated by LSF procedure to the delayed probe stimulus (Long and Talmadge, 1997; Talmadge et al., 1999; Naghibolhosseini et al., 2014). Real time data processing was controlled by a Fireface UC, using Matlab and Psychtoolbox (Kleiner et al., 2007) in a Windows OS programing environment. The SFOAE fine structure obtained was high-pass filtered with a cutoff frequency of 1 cycle/octave to reduce the impact of any slow changes in overall SFOAEs level (Figure 1). The levels of Max and Min frequency were determined by the filtered SFOAE fine structure.

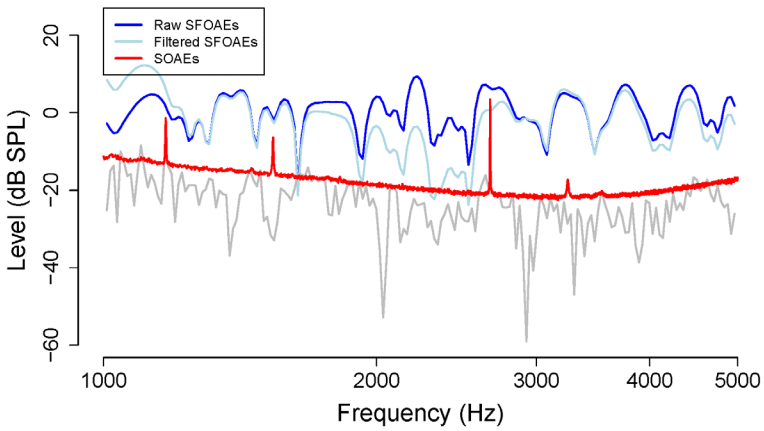


Figure 1. An example of raw (light blue) and the filtered (blue) SFOAEs. The raw SFOAE fine structure (light blue line) was high-pass filtered with a high-pass frequency of 1 cycle/octave to reduce the impact of any slow changes in overall SFOAEs level (blue line) on estimates of fine structure depth. The red line represents SOAEs and the grey line represents noise floor.

2.3 Behavioural task: level discrimination

A two-interval, forced-choice procedure was used: two-tone complexes were presented in two intervals, standard and target, on each trial. All stimuli were presented monaurally at a duration of 300 ms with cosine-squared, 20-ms rise/fall ramps. In the target interval, the level of each tone was always 3 dB greater than that in the standard interval. Small independent and random perturbations in level of the individual tones were presented from one presentation to the next. The level perturbations were normally distributed with sigma=3dB. The order of standard and target intervals was selected at random on each trial. Listeners were asked to choose the interval in which the target (louder) sound occurred by pressing a mouse button. Correct feedback was given immediately after the listener's response. Decision weights on the tones for

each listener were then estimated from logistic regression coefficients in a general linear model for which the perturbations were predictor variables for the listener's trial-by-trial response (Berg, 1990). In experiment 1, the frequencies of the two-tone complex were chosen from SOAE measures from each listener: one at the frequency of an SOAE and the other at a nonSOAE frequency, either lower or higher than the chosen SOAE frequency. The level of each tone in the standard interval was 50 dB SPL. SOAEs usually occur near a maxima of SFOAEs fine structure (Bergevin et al., 2012; Dewey & Dhar, 2014), but SOAEs are not always detectable at such maxima. Thus, we decided to measure SFOAE fine structure and select frequencies at the maxima (Max) and minima (Min) of the fine structure for the behavioural level discrimination task. In experiment 2, two frequencies were chosen from the measured SFOAE fine structure for each listener: one at maxima of the fine structure and the other at a minimum. The level of the standard stimuli was 35 dB SPL During a testing session, SOAEs and SFOAEs were recorded prior to and after the behavioural task.

3 Results

Figure 2 shows the mean relative decision weight for the tone at an SOAE frequency as a function of the mean level-difference of the OAE at the SOAE and non-SOAE frequency for six different ears. There was a statistically significant positive correlation between the relative decision weights and the OAE level difference ($r^2 = 0.759$, $p=0.0148$).

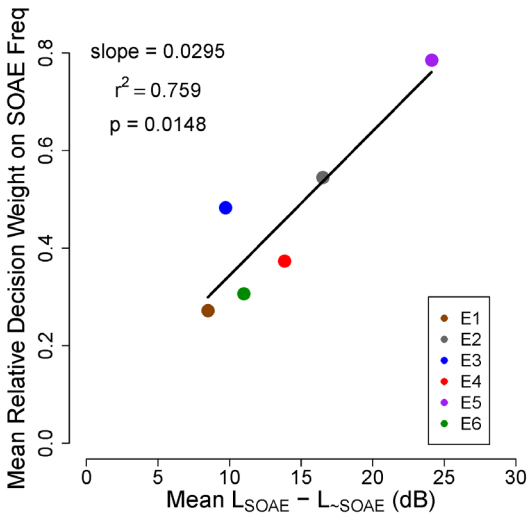


Figure 2. The mean relative decision weight for the tone at an SOAE frequency as a function of the mean OAE level difference measured at the SOAE and nonSOAE (~SOAE) frequency. Each decision weight was based on at least 400 trials. The relative decision weight was calculated as $W_{SOAE\ tone}$ / ($W_{SOAE\ tone}$ + $W_{nonSOAE\ tone}$). Each different colour represents a different ear.

Figure 3 shows the results for comparison of the relative decision weights to the level differences between tones at the maxima and minima of the fine structure for two listeners.These listeners have detectable SOAEs, which occur near SFOAE maxima.

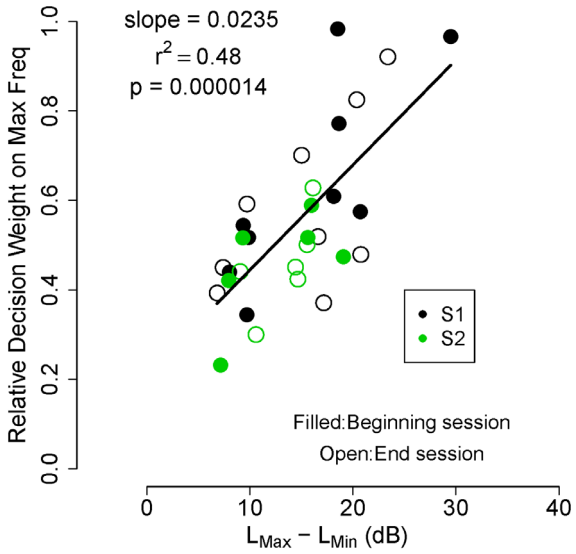


Figure 3. The relative decision weights from two listeners for the tones at SFOAE maxims and minima as a function of the level difference between the maxima and minima. Each decision weight was based on 1000 trials. The relative decision weight was calculated as $W_{Max\ freq} / (W_{Max\ freq} + W_{Min\ freq})$.

The SFOAEs levels obtained at the beginning of the session are associated with the decision weights from the first half of the behavioural trials (filled symbols), and those at the end of session are associated with the decision weights from the second half of the behavioural trials (open symbols). The correlation between the relative decision weight and the level difference between tones near fine-structure maxima and tones near minima is statistically significant (r^2 = 0.48, p=0.000014). This outcome suggests that the association between decision weights and OAEs does not depend on detection of SOAEs.

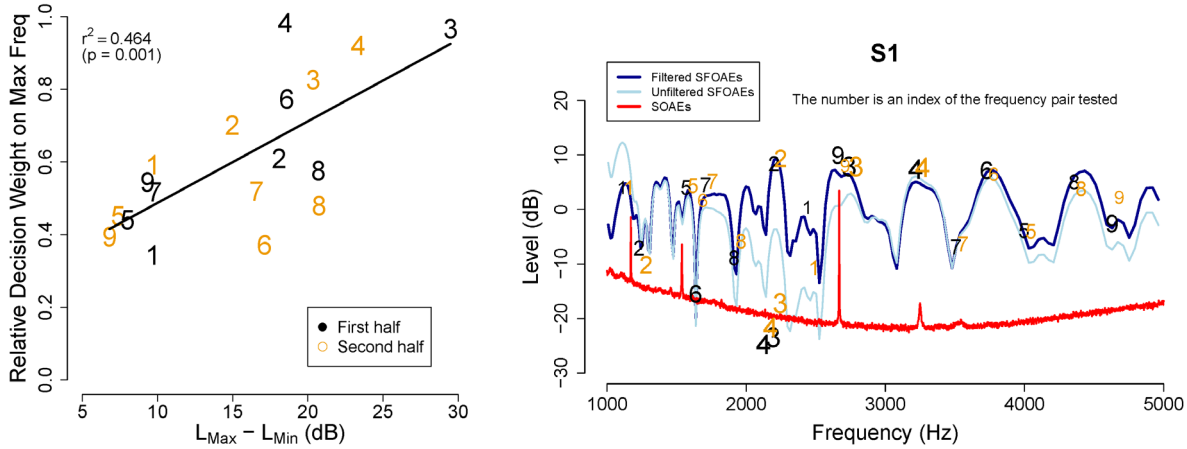


Figure 4. Left panel: The relative weight of the tone at SFOAE fine structure maxima is plotted as a function of the level difference between SFOAE maxima and minima for subject 1. Right panel: The regression curve given in the left panel is used to convert the relative weights to dB differences for comparison to the maxima and minima of the fine structure measured for subject 1. The relative weights as dB differences are identified by number pairs for direct comparison to the relative weights given in the right panel. The size of the numbers represents the relative strength of the decision weights for each frequency pair. Weights from the first half (black) and second half (orange) of the behavioural trials are both presented.

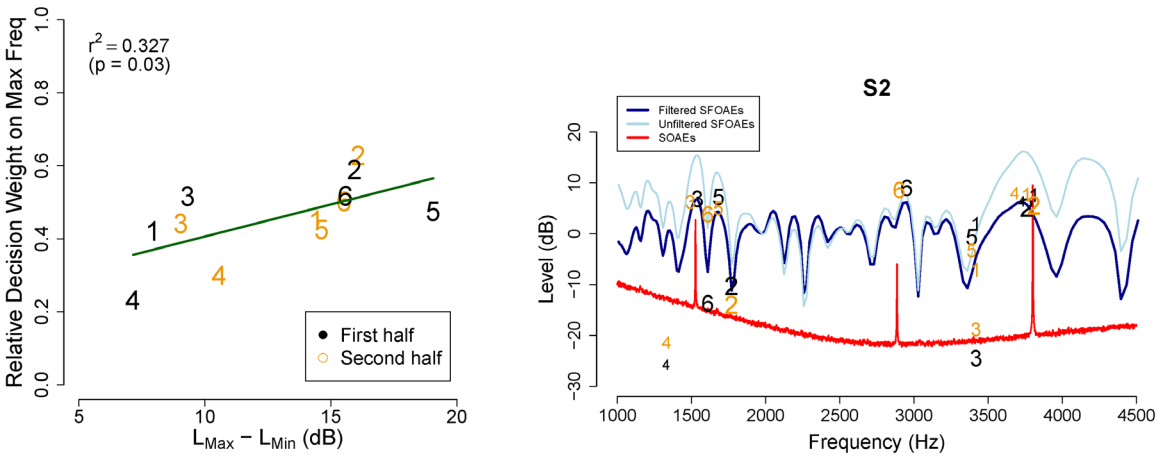


Figure 5. The same as in Figure 4 for subject 2

The position of the frequency pairs used for the behavioural task and the decision weight analysis for subject 1 and 2 are shown in Figure 4 and 5 respectively. The relative weights were transformed into level differences in dB using the fitted function in the right panel and plotted in the right panel at their corresponding SFOAE frequencies.

4 Discussion

Given that loudness varies with SFOAE maxima and minima and that loudness is a strong predictor of listener decision weights, we hypothesized that both SFOAEs and SOAEs may be used to predict individual differences in decision weights in a level discrimination task. As expected, the data showed a significant positive correlation *between* the level difference of the OAE at the SOAE and non-SOAE frequency *and* the relative decision weights on level discrimination task of two-tone complex (Figure 2). Also the same positive correlation was shown with the level difference of maxima and minima of SFOAE's fine structure (Figure 3). For clinical applications swept frequency SFOAEs would likely provide a better measure of cochlear fine structure inasmuch as they are less time consuming than SOAE measurements and provide a clearer indication of regions of threshold microstructure and variation in loudness. These preliminary results suggest that OAEs may ultimately be used as a clinical tool to diagnose individual differences in decision weights in spectral discrimination tasks typical of everyday listening

Acknowledgments

This research was supported by NIDCD grant R01 DC001262-21

References

Alexander, J.M., & Lutfi, R.A. (2004). Informational masking in hearing-impaired and normal-hearing listeners: sensation level and decision weights. *J. Acoust. Soc. Am.*, 116(4), 2234-2247.

Baiduc, R.R., Lee, J., & Dhar, S. (2014). Spontaneous otoacoustic emissions, threshold microstructure, and psychophysical tuning over a wide frequency range in humans. *J. Acoust. Soc. Am.*, 135(1), 300-314.

Berg, B.G. (1990). Observer efficiency and weights in a multiple observation task. *J. Acoust. Soc. Am.*, 88, 149-158.

Bergevin, C., Fulcher, A., Richmond, S., Velenovsky, D., & Lee, J. (2012). Interrelationships between spontaneous and low-level stimulus-frequency otoacoustic emissions in humans. *Hearing Research*, 285, 20-28.

Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and two ears. *J. Acoust. Soc. Am*. 25(5), 975-979.

Dewey, J.B., & Dhar, S. (2014). Comparing behavioral and otoacoustic emission fine structures. *Forum Acousticum*.

Doherty, K.A., & Lutfi, R.A. (1996). Spectral weights for overall discrimination in listeners with sensorineural hearing loss. *J. Acoust. Soc. Am*., 99(2), 1053-1058.

Epstein, M., & Silva, I. (2009). Analysis of parameters for the estimation of loudness from tone-burst otoacoustic emissions. *J. Acoust. Soc. Am*., 125(6), 3855–10. doi:10.1121/1.3106531

Jesteadt, W., Valente, D.L., Joshi, S.N., & Schmid, K.K. (2014). Perceptual weights for loudness judgments of six-tone complexes. *J. Acoust. Soc. Am*., 136(2), 728-735.

Kemp, D.T. (1978). Stimulated otoacoustic emissions from within the human auditory system. *J. Acoust. Soc. Am*. 64(5), 1386-1391.

King, K., Stephens, D. (1992). Auditory and psychological factors in auditory disability with normal hearing. *Scand. Audiol.,* 21, 109-114.

Kleiner M., Brainard, D., & Pelli, D. (2007) "What's new in Psychtoolbox-3." *Perception,* 36.14

Long, G.R., Tubis, A. (1988). Investigations into the nature of the association between threshold microstructure and otoacoustic emissions. *Hearing Research*, 36, 125-138.

Long, G. R., and Talmadge, C. L. (1997). "Spontaneous otoacoustic emission frequency is modulated by heartbeat," *J. Acoust. Soc. Am*., 102, 2831–2848.

Long, G.R., & Talmadge, C.L. (2007). New procedure for measuring stimulus frequency otoacoustic emissions. *J. Acoust. Soc. Am.,* 122, 2969.

Long, G.R., Talmadge, CL, Jeung, C. (2008). New Procedure for evaluating SFOAEs without suppression or vector subtraction, *Association for Research in Otolaryngology*, Vol. 31

Lutfi, R. A. (1995). Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks. *J. Acoust. Soc. Am.,* 97, 1333-1334.

Lutfi R.A., Liu C.J. (2007) Individual differences in source identification from synthesized impact sounds. *J. Acoust. Soc. Am.,* 122, 1017–1028

Lutfi, R.A., Jesteadt, W. (2006). Molecular analysis of the effect of relative tone on multitone pattern discrimination. *J. Acoust. Soc. Am.,* 120(6), 3853-3860.

Lutfi, R.A., Liu, C. J., Stoelinga, C.N.J. (2013). A New approach to sound source segregation. *Basic Aspects of Hearing, Advances in Experimental Medicine and Biology*, 787, 203-211.

Mauermann, M., Long, G.R., Kollmeier, B. (2004). Fine structure of hearing threshold on loudness perception. *J. Acoust. Soc. Am*., 116(2), 1066-1080.

Naghibolhosseini, M., Hajicek, J., Henin S., Long, G.R. (2014), Discrete and Swept-Frequency SFOAE With and Without Suppressor Tones, *Association for Research in Otolaryngology*, 37(73).

Richards, V.M. (2002). Effects of a limited class of nonlinearities on estimates of relative weights. *J. Acoust. Soc. Am*., 111, 1012-1017.

Rasetshwane, D. M., Neely, S. T., Kopun, J. G., & Gorga, M. P. (2013). Relation of distortion-product otoacoustic emission input-output functions to loudness. *J. Acoust. Soc. Am*., 134(1), 369–15. doi:10.1121/1.4807560

Talmadge, C. L., Long, G. R., Tubis, A., and Dhar, S. (1999). "Experimental confirmation of the two-source interference model for the fine structure of distortion product otoacoustic emissions," *J. Acoust. Soc. Am*., 105, 275–292

Thorson, M. J., Kopun, J. G., Neely, S. T., Tan, H., & Gorga, M. P. (2012). Reliability of distortion-product otoacoustic emissions and their relation to loudness. *J. Acoust. Soc. Am.*, 131(2), 1282–14. doi:10.1121/1.3672654

Chapter 50

On the Interplay Between Cochlear Gain Loss and Temporal Envelope Coding Deficits

Sarah Verhulst, Patrycja Piktel, Anoop Jagadeesh, Manfred Mauermann

Medizinische Physik and Cluster of Excellence Hearing4all, Dept. of Medical Physics and Acoustics, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Contact author: SV, sarah.verhulst@uni-oldenburg.de

Abstract

Hearing impairment is characterized by two components: (i) cochlear gain loss that yields wider auditory filters and elevated hearing thresholds, and (ii) cochlear neuropathy, a noise induced aspect of hearing loss that may impact temporal coding fidelity of supra-threshold sound. This study uses a psychoacoustic amplitude modulation (AM) detection task in quiet and multiple noise backgrounds to test whether these aspects of hearing loss can be isolated in listeners with normal to mildly impaired hearing ability. Psychoacoustic results were correlated to hearing threshold, categorical loudness scaling, distortion product otoacoustic emission (DPOAE) thresholds and envelope-following response (EFR) measures. AM thresholds to noise carriers (4 kHz octave-wide) depended on DPOAE thresholds and associated cochlear compression levels. The amount with which wideband masking noise (1 octave) worsened 4 kHz pure-tone AM thresholds was inversely correlated to DPOAE thresholds in correspondence to a compression loss benefit, while the amounts with which narrowband noise maskers (40 Hz) worsened performance was not. While the latter condition was designed to reduce temporal coding ability within an auditory filter, the correlation with the EFR measures was weak. This study elucidates how psychoacoustic difference measures can be adopted in the differential diagnostics of hearing deficits in listeners with mixed pathologies.

Keywords: Hearing Impairment Diagnostics, Cochlear Neuropathy, Amplitude Modulation Detection, EFR, DPOAE

1 Introduction

Listeners with impaired audiograms likely suffer from a combination of pathologies that may interact to affect their speech intelligibility in adverse listening conditions. The well-studied cochlear gain loss aspect of hearing loss due to outer-hair-cell deficits is known to impact audibility of sound, whereas cochlear neuropathy reduces the number and types of auditory nerve (AN) fibers responsible for robust afferent transmission of sound (Kujawa and Liberman, 2009). Cochlear neuropathy is thought to affect sound encoding at supra-threshold levels as high-threshold low-spontaneous rate (LS) AN fibers are most sensitive to noise exposure (Furman et al., 2013). Indeed, a recent study has demonstrated that listeners with normal hearing thresholds can show supra-threshold hearing deficits (e.g., in envelope ITD, AM detection threshold tasks) that are related to temporal coding fidelity in the auditory brainstem while being uncorrelated to audiometric or DPOAE thresholds (Bharadwaj et al., 2015).

Even though a temporal coding deficit may influence auditory perception, it is currently not known how cochlear neuropathy interacts with the cochlear gain loss aspect of hearing loss, or whether it is equally important for auditory perception. On one hand AM detection is expected to *improve* when cochlear compression is reduced (Moore et al., 1996), while cochlear neuropathy may *degrade* temporal coding fidelity (Bharadwaj et al., 2014, 2015). To study the interaction between these components, and to quantify contributions in listeners that may suffer from both aspects of hearing loss, we tested amplitude-modulation (100 Hz) detection in quiet and in the presence of noise maskers. To force the system to rely on redundancy of coding within a single auditory filter, we determined pure-tone AM detection thresholds in a fixed-level narrowband noise masker (NB; 40 Hz) condition. The difference between the AM threshold in quiet and with the NB noise masker might be a metric that is free from cochlear compression and sensitive to the temporal coding fidelity aspect of hearing loss. A second difference measure using a fixed level wideband noise masker (BB, 1 octave) tests whether cochlear neuropathy is equally important than auditory filter width in processing temporal modulations in the presence of background noise.

We tested 21 normal and mildly hearing-impaired listeners and correlated psychoacoustic AM thresholds to their hearing and DPOAE thresholds. Cochlear compression was estimated from the compressive range of DPOAE growth curves and the lower slope of the categorical loudness-scaling curve. Temporal coding ability was assessed using EFR strength to 100% modulated broadband noise, and to an EFR slope measure that tests how quickly EFRs degrade as a function of modulation depth reduction.

2 Methods

We selected listeners in two categories: either having a normal 4 kHz pure-tone threshold, or having a slight hearing loss (25 dB HL) at 4 kHz. The test population was formed with 10 subjects (4 male, 6 female), aged from 20 to 32 (mean: 25.9), that had near normal hearing thresholds (<15 dB HL, flat), 4 subjects (3 male, 1 female) aged from 25 to 35 (mean: 25.9), with a normal 4 kHz threshold, but a slightly sloping audiogram in the higher frequencies, and 7 subjects (3 male, 4 female) aged from 39 to 75 (mean: 60.7), with a near 25 dB dB loss at 4 kHz and a mild hearing loss at the higher frequencies. The populations are indicated with different colors throughout the plots as follows: normal hearing (blue), very slight (red), mild hearing loss (black).

Sound delivery for OAEs, EFRs, and AM detection threshold measurements was provided by ER-2 insert earphones attached to a TDT-HB7 headphone driver and Fireface UCX soundcard. Stimuli were generated in Matlab and calibrated using a B&K type 4157 ear simulator and sound level meter. Audiograms were measured using a HDA 200 headphone and an AT900 Auritec audiometer. Loudness scaling was measured using the ALCALOS Procedure (Brand and Hohmann, 2002) using the OMA software package (HörTech) with the standard 1/3rd octave bands of low-noise noise presented over HDA 200 headphones. OAEs were recorded using the OLAMP software and an ER10B+ microphone, EFRs were recorded using a 32-channel Biosemi EEG amplifier using a custom built triggerbox, and analyzed using the ANLFFR and Matlab software.

2.1 Amplitude Modulation Detection Thresholds

Amplitude-modulation detection thresholds for 4-kHz, 500-ms pure-tone and 1 octave-wide noise carriers were obtained using a 3AFC method (1-up, 2 down) in quiet and in masking noise. Thresholds were obtained from the last six reversals at the smallest step size and 4 repetitions were measured of which the first run was discarded. The initial modulation depth value was -6dB ($20 \cdot \log_{10}(m)$, m=50%), and varied adaptively with stepsizes of 10, 5, 3 and 1 dB. The modulation frequency was 100 Hz to target brainstem auditory processing (Purcell et al., 2004),

while providing the modulation within a single equivalent rectangular bandwidth (ERB). Target levels were 70 dB SPL for the pure tone and NH listeners, and adjusted to the level corresponding 25 categorical units in the loudness scaling (approx. 45-50 dB SL) to allow for equal sensation in the different subject groups. The noise carrier was presented at an rms level of 25 dB SPL lower than that of the pure tone carrier (i.e., 45 dB SPL for the NH listeners). Low-level 2-octave wide masking noise (Target level – 30 dB) was present in the noise carrier condition to avoid off-frequency cues (Moore et al., 1992). Such a low level masking noise was not present in the pure-tone carrier condition. The two on-frequency maskers to study the influence of coding fidelity (narrow-band noise; NB) and auditory filter width (broadband noise; BB) were presented at levels of target – 15 dB spectral level (NB) and target – 50 dB spectral level (BB) and had bandwidths of centered around 4 kHz of 40 Hz (NB) and 1 octave (BB).

2.2 Envelope Following Responses

EFRs were obtained for wideband noise carriers presented at 70 dB SPL, irrespective of hearing threshold for a modulation frequency of 120 Hz and modulation depths ($20\log_{10}(m)$) of -8, -4 and 0 dB relative to m of 100%. 600 repetitions of 600 ms modulated epochs were recorded on 32 channels, and after filtering (60-600 Hz), artefact rejection (100 µV), epoching and baseline correction, the FFT of the averaged epochs in each of the 32 channels was calculated. EFR strength (in dB) was determined as the spectral level at the frequency of the modulator and averaged across the 32 channels. Both the EFR strength of the 100% modulated condition was considered, and the slope of the EFR strength as a function of modulation depth reduction as the later measure reflects temporal coding fidelity (Bharadwaj et al, 2015). Slopes were only considered when at least the 0 and -4 dB EFR levels were above the noise floor; if the -8 dB EFR was not present, the level of the noise floor was used as the level of the -8 dB EFR strength. The slope was thus calculated using a linear fit to 3 datapoints: 0, -4 and -8 dB corresponding to the modulation depths tested.

2.3 DPOAE Thresholds and Compression

Distortion product otoacoustic emission (DPOAE) based threshold and compression indicators were derived from $2f_1$-f_2 DPOAE I/O measurements using the DPOAE sweep method (Long et al., 2008). The primary frequencies were exponentially swept up (2s/octave) over an 1/3 octave range around the geometric mean of 4 kHz at a constant frequency ratio f_2/f_1 of 1.2. Using a sufficiently sharp least squared fit filter (here ca. 2.2 Hz), the distortion component (DCOAE) can be extracted from the DPOAE recording (Long et al., 2008). The DCOAE is generated around the characteristic site of f_2 and thus predominantly provides information about the f_2 site without being influenced by DPOAE fine structure that is known to affect I/O functions unwantedly (Mauermann and Kollmeier, 2004). DCOAE I/O functions were computed as average over 34 DCOAE I/O functions across the measured frequency range. A matched cubic function $L_{DP} = a + \left(\frac{1}{q} \cdot (L_2 - b)^3\right)$ with parameters a, b, and q was fit to the data points. All fits yielded r^2 values of 0.98 or above. DPOAE thresholds were determined as the level of L_2 at which the extrapolated fitting curve would reach a level of -25 dB SPL (~0 Pa). As additional indicator for cochlear compression we used the intermediate compressive range, which is the difference between the L_2 levels at which the fitting function has a slope of 0.4 and 0.75 dB/dB.

3 Results

Figure 1 depicts correlations between the hearing threshold (dB HL), the lower slope of the loudness scaling function (C.U./dB) and the EFR measures. A significant correlation (p=0.02) between EFR strength at 100% modulation and hearing threshold was found. Despite the large spread of the data in the vertical direction, this correlation might have occurred because the broadband EFR stimuli were not adjusted to the specific hearing loss configuration. Our results

did not seem to show a compression loss benefit for AM coding in which contradicts to the slightly enhanced EFRs recorded in chinchillae with sensorineural hearing loss when stimuli were presented relative to the hearing threshold (Zhong et al., 2014). Importantly, the spread of EFR strength across listeners with similar hearing thresholds in Figure 1A is large demonstrating large individual differences in EFR strength for equal hearing thresholds. Because EFR strength did not correlate ($p>0.75$) to the lower loudness curve slope, nor to the DPOAE threshold and compression measure, it appears that EFR strength may also contain a component of scalp size, or temporal coding fidelity influenced by either the absolute number or types of AN fibers contributing to the population response. DPOAE thresholds were linked to the lower loudness curve slope and DPOAE compression ($p<0.01$), but were unrelated to EFR strength ($p=0.7$).

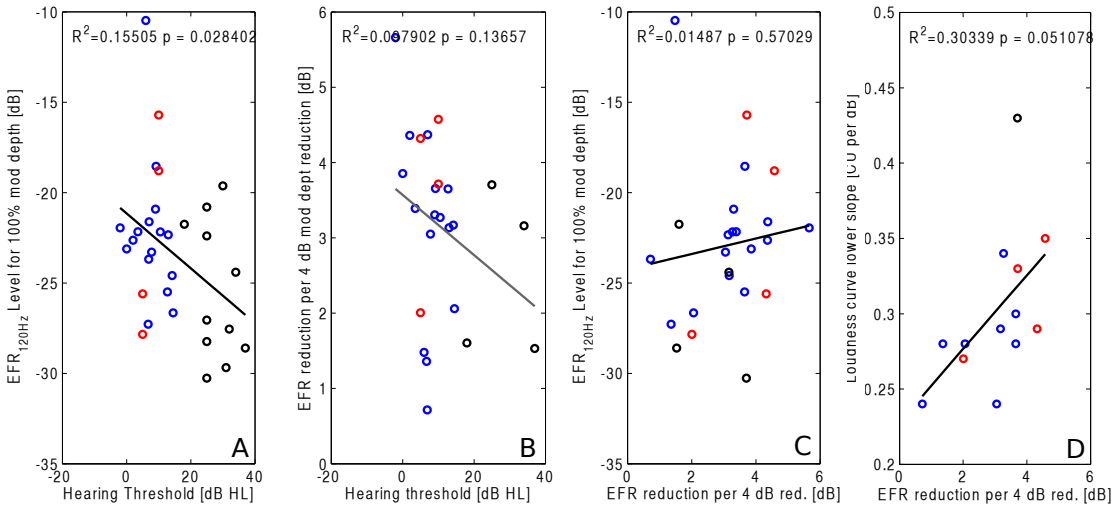


Figure 1. Correlations between Hearing Threshold and EFR strength measures (A and B) and EFR slopes as a function of modulation depth reduction and EFR strength (C) and Loudness scaling slopes (D). Data from 10 additional NH listeners participating in another study (Verhulst et al., 2015) were added in panels A-C.

The EFR slope measure that reflects how quickly EFRs reduce AM coding strength as a function of modulation depth reduction was less influenced by hearing thresholds (panel B), and did not correlate to EFR strength (panel C). The latter observation confirms that the EFR slope metric reflects another aspect of sound encoding that may link to temporal coding fidelity of supra-threshold sound, as suggested by Bharadwaj et al. (2015). The near correlation between the EFR slope metric and loudness slope measure indicates that either (i) compression loss does have an influence on this metric, or that (ii) listeners with steeper loudness curves also exhibit reduced temporal coding ability. The finding that the EFR slope measure was uncorrelated to DPOAE thresholds ($p=0.44$) and compression ($p=0.5$) suggests the latter.

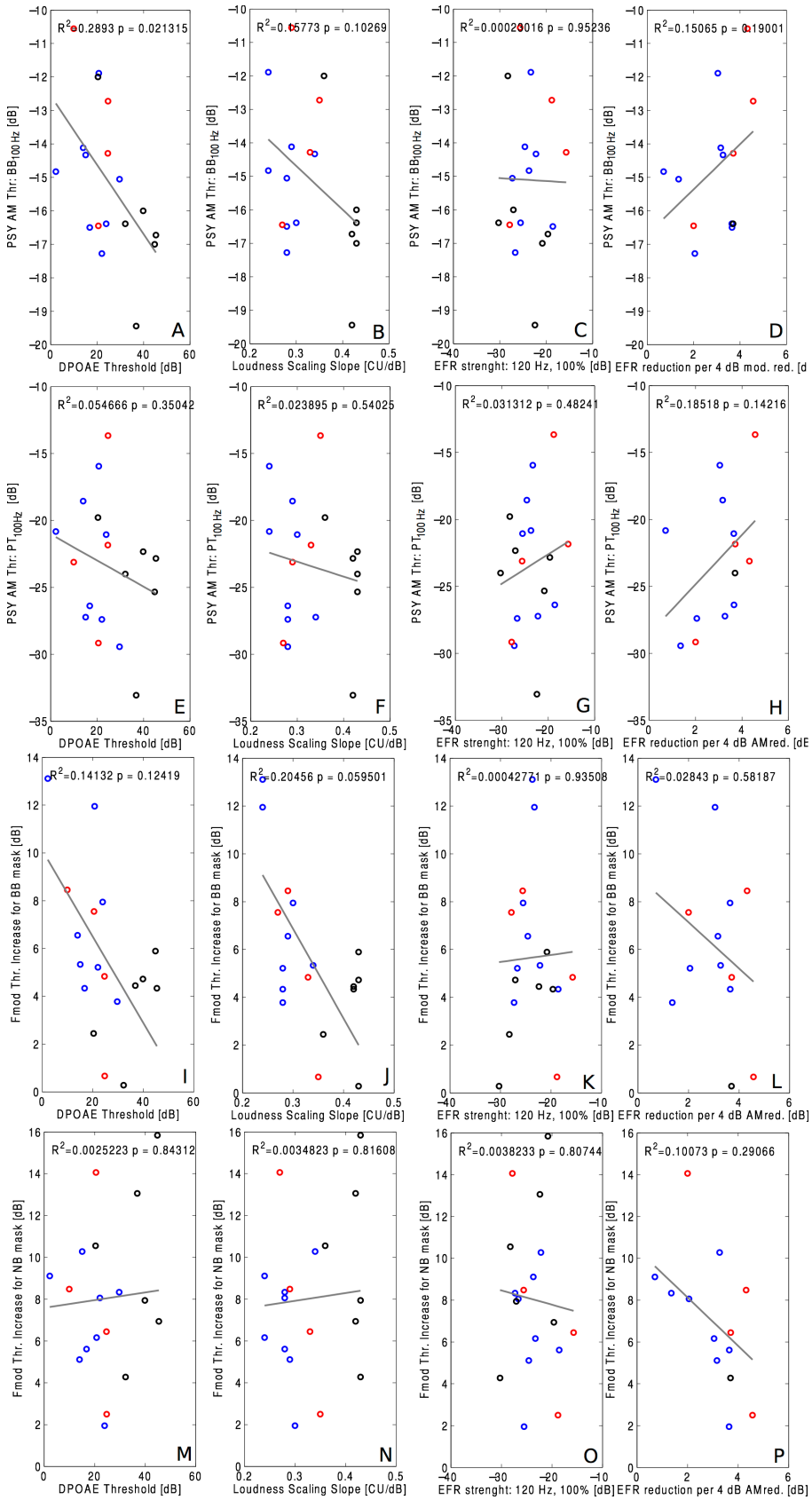


Figure 2. Correlations between psychoacoustic AM detection thresholds and DPOAE, loudness and EFR measures. Panels A-D: 1-oct broadband carrier, E-H pure-tone carrier, I-L: Modulation detection difference between the pure-tone carrier in 1-octave-wide (carrier-30 dB) masking noise and the carrier in quiet. M-P: similar for the tone carrier in 40 Hz wide (carrier-15 dB) masking noise.

The correlations in Figure 2 show the relationship between amplitude modulation detection in quiet (top 2 rows) and in the presence of background noise (bottom 2 rows) to study the interactions between cochlear gain loss and temporal coding deficits. AM detection for the 1-octave-wide noise carrier (A-E) has the closest correspondence to the wide-band stimuli used for the EFR measures, and showed that listeners with higher DPOAE thresholds at 4 kHz had better detection thresholds (A; $p<0.05$). This correlation was weaker for the pure-tone carrier condition (E), but the relationship for the NH people stayed similar. Because DPOAE thresholds correlated significantly to DPOAE compression and the lower loudness curve slope, outer-hair-cell related gain loss seems to improve amplitude modulation detection in quiet The latter result is in agreement with findings and explanations proposed in the Moore et al. (1996) and Stone and Moore (2014) studies.

It is expected that adding a noise masker that is wider than the ERB would cause two effects: (i) it would increase the amount of noise within the auditory filter, where wider auditory filters would pass through more noise, and (ii) it would make reliable coding of AM information within an ERB more noisy in each coding channel (e.g., in each auditory nerve fiber). We expect that subjects with neuropathy are more affected by the noise masker because they have fewer coding channels available. Secondly, even though compression loss may improve AM encoding in quiet, the total amount of noise that falls within a single auditory filter could counteract the compression loss benefit for AM detection.

Panels I-L show that the amount with which AM detection thresholds were elevated by the presence of the BB noise affected listeners with low DPOAE thresholds more so than listeners with elevated DPOAE thresholds. Note also that the loudness scaling slope measure correlated both to the EFR slope measure ($p=0.05$) and DPOAE threshold ($p<0.001$) measures. The relation between AM thresholds and the DPOAE and loudness slope measures suggest that listeners with compression loss, benefit in the presented background noise despite the fact that we can assume that those listeners would have wider auditory filters yielding reduced SNRs in a single ERB, and may also exhibit a temporal coding deficit. The concomitant weak correlation with the EFR slope measure suggests that cochlear compression may be the more confounding factor in modulation detection in the presence of fixed a signal-to-noise background noise.

Lastly, the NB masker condition was designed to affect temporal coding fidelity within a single auditory filter (M-P). The amount with which the AM detection threshold is elevated in response to the added NB noise masker should inform about temporal coding deficits without being influenced by the width of the auditory filter in itself. In line with that idea, the weak trends between the DPOAE and loudness scaling slope measures from the quiet condition (E-F) have disappeared from the NB masker conditions (M-N). As the EFR slope condition is expected to link to the temporal coding fidelity aspect of AM coding, a strong relationship between this measure and the AM threshold elevation in the NB masker is expected. Unfortunately, this relationship was not observed; rather an inverse trend to the quiet condition was observed as a result from the lacking relationship between the AM threshold in the NB noise condition and the EFR slope measure ($p=0.9$). Thus, even though the relative metric is free from DPOAE threshold effects, the narrowband noise did not have the envisioned relation with the EFR slope measure. There are three factors that could be at the basis this observation: (i) perhaps the NB noise showed too many temporal envelope fluctuations such that it did not only add noise to the individual coding channels, but introduced an envelope cue that interfered with AM detection. (ii) The EFR stimulus was broadband and thus provided an estimate of temporal coding fidelity across the cochlea; it is worthwhile testing whether the relationship improves for EFRs to a pure-tone carrier. (iii) The hearing threshold compensation that was used in the psychoacoustic measures was not applied to the EFR measures, which means that at the frequency of interest, the modulated 100% EFR could have been processed at a different operating point than was the psychoacoustic stimulus. Even though the stimulus level compensation in the loudness criterion is much softer than, e.g., a correction to dB SL would have been, it might have ended up shifting the input for the hearing impaired listeners into the low spontaneous rate fibre coding regime

such that hearing impaired listeners with no neuropathy would benefit from the more precise coding in those level ranges for the psychoacoustic tasks whereas this benefit not present in the EFR measures.

4 Discussion

The present study investigated how cochlear neuropathy and cochlear gain loss interact to affect perception of fast temporal modulations important for speech perception in adverse conditions. While both hearing and DPOAE thresholds may influence the 100% modulation depth conditions (Figure 2 rows 1 and 2), the psychoacoustic difference measures were designed to show the detrimental effect a specific noise masker had on a listeners' performance. While the wide-band noise condition seemed to affect listeners with smaller DPOAE thresholds and reduced compression more, this effect was gone for the narrowband noise condition indicating that the psychoacoustic difference measure was successful in bypassing the cochlear gain loss influence in the narrowband difference measure. However, the narrowband noise difference measure did not confirm our temporal coding deficit hypothesis. Given that we only studied a small group of listeners, and that the EFR stimulus was broadband while the psychoacoustic stimulus was narrowband, it is worthwhile pursuing whether this correlation between the temporal coding EFR metric and the perception can be improved. To conclude, differential psychoacoustic methods form a promising candidate to separate different aspects of hearing loss in listeners with mixed pathologies.

Acknowledgments

The authors thank Frauke Ernst and Anita Gorges for assistance with EEG measurements and subject recruitment and Steven van de Par for discussions related to this work. Work supported by the DFG Cluster of Excellence EXC 1077/1 "Hearing4all"

References

Bharadwaj, H., Verhulst, S., Shaheen, L., Liberman, M.C., & Shinn-Cunningham, B. (2014). Cochlear Neuropathy and the coding of supra-threshold sound. *Frontiers in Systems Neuroscience*, 8, 26.

Bharadwaj, H., Masud, S., Mehraei, G., Verhulst, S., & Shinn-Cunningham, B. (2015). Individual Differences Reveal correlates of hidden hearing deficits. *Journal of Neuroscience,* 35 (5), 2161-2172.

Brand, T. & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *Journal of the Acoustical Society of America,* 112 (4), 1597.

Furman, A.C., Kujawa, S.G., & Liberman, M.C. (2013). Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *Journal of Neurophysiology,* 110(3), 577.

Kujawa, S.G., & Liberman, M.C. (2009). Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss. *Journal of Neuroscience,* 29, 14077.

Long, G.R., Talmadge, C.L., & Lee, J. (2008). Measuring distortion product otoacoustic emissions using continuously sweeping primaries, *Journal of the Acoustical Society of America*, 124, 1613–1626.

Mauermann, M., & Kollmeier, B. (2004). Distortion product otoacoustic emission (DPOAE) input/output functions and the influence of the second DPOAE source, *Journal of the Acoustical Society of America*, 116, 2199.

Moore, B.C., Shailer, M.J., & Schooneveldt, G.P. (1992). Temporal modulation transfer functions for band-limited noise in subjects with cochlear hearing loss. *British journal of audiology*, *26*(4), 229-237.

Moore, B.C., Wojtczak, M., and Vickers, D.A. (1996). Effect of loudness recruitment on the perception of amplitude modulation. *Journal of the Acoustical Society of America*, 100, 481–489.

Purcell, D.W., John, S.M., Schneider, B.A., & Picton, T.W. (2004). Human temporal auditory acuity as assessed by envelope following responses. *Journal of the Acoustical Society of America.*, 116(6), 3581.

Stone, M.A., & Moore, B.C. (2014). Amplitude-modulation detection by recreational-noise-exposed humans with near-normal hearing thresholds and its medium-term progression. *Hearing Research*, *317*, 50-62.

Verhulst, S., Ernst, F., van de Par, S. (2015). In Search for the cochlear Neuropathy Component of Audiometric Hearing loss, *Association for Research in Otolaryngology Midwinter Meeting,* Poster.

Zhong, Z., Henry, K.S., Heinz, M.G. (2014) Sensorineural hearing loss amplifies neural coding of envelope information in the central auditory system of chinchillas. *Hearing Research,* 309, 55-62.

Chapter 51

Frequency Tuning of the Efferent Effect on Cochlear Gain in Humans

Vit Drga
Ear Institute, University College London (UCL), London, UK

Christopher J. Plack
School of Psychological Sciences, The University of Manchester, Manchester, UK

Ifat Yasin
Ear Institute, University College London (UCL), London, UK

Contact author: V.D. Email address: v.drga@ucl.ac.uk

Abstract

Cochlear gain reduction via efferent feedback from the medial olivocochlear bundle (MOCB) is frequency specific (Guinan, 2010). The present study with humans used the Fixed Duration Masking Curve (FDMC) psychoacoustical method (Yasin et al., 2013a,b, 2014) to estimate the frequency specificity of the efferent effect at the cochlear level. The combined duration of the masker-plus-signal stimulus was 25 ms, within the efferent onset delay of about 31-43 ms (James et al., 2002). Masker level (4.0 or 1.8 kHz) at threshold was obtained for a 4-kHz signal in the absence or presence of an ipsilateral 60-dB SPL, 160-ms precursor (200-Hz bandwidth) centred at frequencies between 2.5 and 5.5 kHz. Efferent-mediated cochlear gain reduction was greatest for precursors with frequencies the same as, or close to that of, the signal (gain was reduced by about 20 dB), and least for precursors with frequencies well removed from that of the signal (gain remained at around 40 dB). The tuning of the efferent effect filter (tuning extending 0.5 to 0.7 octaves above and below the signal frequency) is within the range obtained in humans using otoacoustic emissions (OAEs) (Lilaonitkul and Guinan, 2009; Zhao and Dhar, 2012). The 10-dB bandwidth of the efferent-effect filter at 4000 Hz was about 1100 Hz (Q_{10} of 3.5). The FDMC method can be used to provide an unbiased measure of the bandwidth of the efferent effect filter using ipsilateral efferent stimulation.

Keywords: fixed-duration masking curves, forward masking, compression, gain, efferent, precursor, frequency-range

1 Introduction

In addition to ascending (afferent) neural pathways, the mammalian auditory system contains descending (efferent) neural projections from higher to lower levels of the auditory system [ipsilateral and contralateral efferent systems of cat (Liberman, 1988; Huffman & Henson, 1990) and human (Guinan, 2006)]. The mammalian auditory system includes a brainstem-mediated efferent pathway from the superior olivary complex by way of the medial olivocochlear bundle (MOCB) which reduces the cochlear gain applied over time to the basilar membrane (BM) response to sound [electrical stimulation of the olivocochlear bundle (OCB) in guinea-pigs

(Murugasu & Russell, 1996)]. The human MOCB response (as measured using otoacoustic emissions) has an onset delay of between 25 and 40 ms and rise and decay time constants in the region of 280 and 160 ms, respectively (Backus & Guinan, 2006). Recordings from guinea pigs show the greatest reduction in BM vibration due to efferent stimulation when the stimulating tone is close to or above the characteristic frequency (CF) associated with the recording site (Russell & Murugasu, 1997). Similarly, suppression of human otoacoustic emissions (OAEs) via efferent activation is more effective for stimulating frequencies slightly above and below the test probe frequency (Maison et al., 2000).

The advantage of psychoacoustical methods to infer the efferent effect is that they can be used in cases of both normal and mildly impaired hearing (unlike OAEs which can be eliminated by a mild hearing loss). Psychoacoustical studies suggest the efferent effect appears to decrease as the precursor frequency is set higher or lower in frequency than the subsequently presented masker (Bacon & Viemeister, 1985; Bacon & Moore). The present study uses a psychoacoustical forward-masking technique [fixed-duration masking curve (FDMC) method (Yasin et al., 2013a,b, 2014)] in which a precursor sound is presented to elicit the efferent response. The FDMC method avoids the confounds of previous forward-masking methods used to measure the efferent effect by using a combined duration of the masker-plus-signal stimulus of 25 ms, which is within the efferent onset delay. Hence, the representation of the signal should be unaffected by efferent activity elicited by the masker. The effect of efferent activation on cochlear gain can then be separately studied by presenting a precursor sound (to activate the efferent system) prior to the onset of the combined FDMC masker-signal stimulus. The present study used the FDMC method to estimate the frequency range of efferent-mediated cochlear gain reduction in humans.

2 Method

2.1 Listeners

Three normal-hearing listeners participated (L1 (author), L2 and L3). All listeners had absolute thresholds better than 20 dB HL for sound frequencies between 0.25 and 8 kHz (ANSI, 1996) and were practiced in related psychoacoustical experiments. Listeners undertook between 4 and 10 hours of training on subsets of the stimuli used for the main experiment. L2 and L3 were paid £6 per hour of participation.

2.2 Stimuli

The signal was a 4-kHz sinusoid and the sinusoidal masker was either on-frequency (4 kHz) or off-frequency (1.8 kHz); both signal and masker always began in sine phase (Yasin et al., 2013a,b, 2014). The signal had a total duration of 6-ms (3-ms onset and offset ramps, 0-ms steady-state). Absolute thresholds for this signal, presented to the left ear, were 18.2, 20.1 and 23.5 dB SPL for listeners L1, L2 and L3. The masker had a total duration of 19 ms (2-ms onset and offset ramps, 15-ms steady-state). The frequency range of the efferent effect on cochlear gain was studied by presenting a bandlimited precursor sound centred at different frequencies prior to the onset of the combined FDMC masker-signal stimulus (see Figure 1). The ipsilateral precursor, when present, was a 160-ms (5-ms onset and offset ramps, 150-ms steady-state), 200-Hz-wide (brickwall filtered) noiseband presented at 60 dB SPL. The precursor was presented at centre frequencies of 2.5, 3.0, 3.5, 3.75, 4.0, 4.25, 4.5, 5.0 or 5.5 kHz at a precursor offset-masker onset silent interval of 10 ms. A silent interval of 10 ms was chosen to reduce potential confusion effects when the precursor frequency was close to the signal frequency. Stimuli were presented via the left channel of a pair of Sennheiser HD 600 headphones.

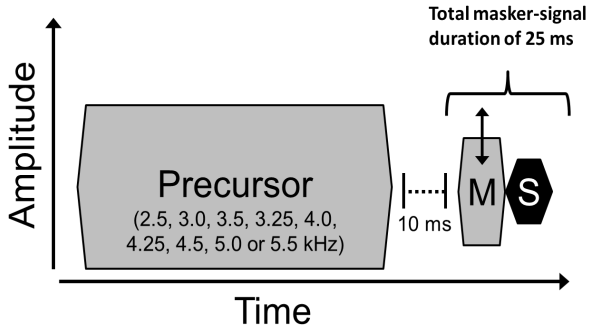


Figure 1. Schematic of the stimuli used in phase 2 of the study. Masker (M) level at threshold for the signal (S) was measured using the FDMC method. Off-frequency or on-frequency masker-plus-signal total duration was 25 ms and masker-signal temporal interval was 0 ms. A precursor was presented prior to the combined masker-signal stimulus at frequencies of 2.5, 3.0, 3.5, 3.75, 4.0, 4.25, 4.5, 5.0 or 5.5 kHz at a level of 60 dB SPL. The silent temporal interval between precursor offset and masker onset was 10 ms. Double-headed arrow indicates that the masker level was adaptively varied in each stimulus trial.

2.3 Procedure

The experiment consisted of two phases (Yasin et al., 2014). In phase 1, for each listener, signal level at threshold was measured per precursor frequency in the presence of a precursor and absence of any tonal masker, to take the forward masking of the precursor into account. In phase 2, the signal level was set at 15 dB SL (based on phase 1 results), a tonal masker was presented as well as the precursor, and masker level at threshold was measured for each precursor frequency and masker frequency. Phase 1 and phase 2 thresholds were also measured for a no-precursor condition. Thresholds were measured using a two-interval, two-alternative forced choice adaptive tracking procedure using a 2-down 1-up rule (phase 1), or a using a 2-up 1-down rule (phase 2) (Levitt, 1971). Trial-by-trial feedback was provided. Per block of trials, the initial adaptive track step size was 5 dB, which reduced to 2 dB after four reversals. Threshold was obtained by averaging stimulus levels for the next eight reversals, but the block was rerun if the standard deviation was greater than 6 dB, or the nominal masker level reached 108 dB SPL or higher (due to soundcard output clipping). Listeners ran in 2-hour sessions, taking breaks as needed. Condition order was randomized within sessions. Reported listener threshold values for phase 2 are the mean of 3-6 separate threshold estimates.

3 Results

The mean data obtained in the presence of a precursor are shown in Figure 3. Masker levels at threshold obtained using a 1.8-kHz masker range from about 60 to 80 dB SPL, and remain roughly constant within this range as precursor frequency is increased from 2.5 to 5.5 kHz. Variability in data from listener L3 accounts for most of the dip seen in this data series. Masker levels at threshold obtained using a 4-kHz masker range from about 25 to 45 dB SPL, with the greatest masker level at threshold (about 45 dB SPL) associated with the presence of a precursor with a frequency close to that of the signal. Comparing masker levels at threshold for on- and off-frequency maskers in the presence of a precursor (elicits efferent activation) with masker levels obtained in the absence of precursor (absence of efferent activation), it can be seen that the greatest change in masker level at threshold (of about 20 dB) occurs when there is an on-frequency masker, rather than an off-frequency masker, and this change is greatest when the precursor frequency is the same as or just below the 4-kHz signal.

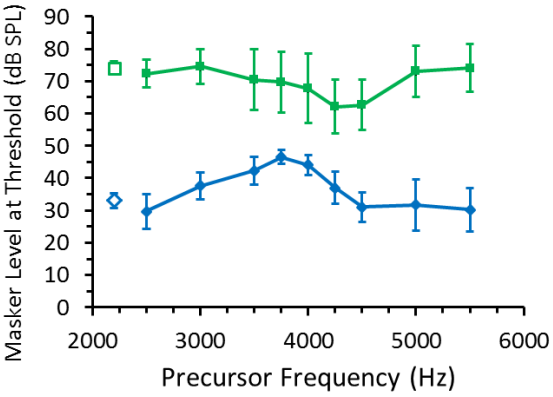


Figure 2. Mean masker level at threshold (across three listeners) as a function of precursor frequency for on-frequency maskers (4 kHz, lower series) or off-frequency maskers (1.8 kHz, upper series). Open symbols at left show masker level at threshold in the absence of a precursor. Error bars show ±1 standard error.

An estimate of inferred basilar-membrane gain can be obtained from the data as the difference between on- and off-frequency masker level values. The mean gain estimated in the absence of a precursor is 41 dB. The estimated gain (as difference between on- and off-frequency masker levels) in the presence of a precursor of different frequencies is shown in Figure 3. Maximum gain estimates of about 40 dB indicate little effect of any efferent-mediated cochlear gain reduction, and are associated with precursor frequencies well above and below that of the signal frequency of 4 kHz. Minimum gain estimates of about 20 dB indicate maximal effect of efferent-mediated cochlear gain reduction, and are associated with precursor frequencies close to that of signal frequency of 4 kHz. The estimated 10-dB bandwidth of the efferent-effect filter is 1100 Hz (Q_{10} of 3.5).

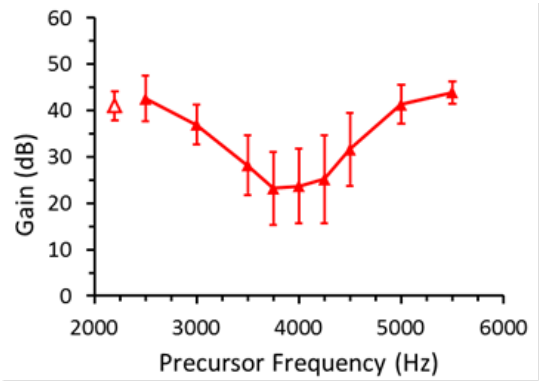


Figure 3. Mean gain estimates (across three listeners) as a function of precursor frequency. Gain is calculated as the difference between masker levels at threshold obtained for on-frequency (4-kHz) and off-frequency (1.8-kHz) maskers (Fig. 2). Open symbol at left shows mean gain value in the absence of a precursor. Error bars show ±1 standard error.

4 Discussion

The FDMC method was used to obtain an estimate of the efferent effect in the absence (without a precursor) and presence (with a precursor presented at different frequencies) of efferent activation to investigate the frequency specificity of the efferent effect on cochlear gain. There was minimum cochlear gain reduction via efferent-activation when the precursor frequency was well removed from that of the signal frequency of 4-kHz, and maximal cochlear gain reduction (about 20 dB) when the precursor frequency was the same as or close to that of the signal. The results suggest that the efferent effect extends for precursors up to about 0.5 to 0.7 octaves above and below the signal frequency. The frequency range of the efferent effect found in the present study is similar to that observed with the measurement of otoacoustic emissions (OAEs) in humans, using simultaneous or contralateral elicitors. The tuning of the efferent effect can extend 0.5-2 octaves above and below the frequency range of interest (spontaneous OAEs) or probe frequency (stimulus-frequency OAEs), as measured using simultaneous ipsilateral/contralateral tones (Lilaonitkul and Guinan, 2009; Zhao and Dhar, 2012). In most cases the tone/narrowband noise eliciting the efferent effect has maximal effect above a presentation level of 60 dB SPL, and in some cases the efferent effect has been shown to be more

pronounced if the elicitor/precursor sound is about 0.5-1 octave below the probe frequency (e.g., Mott et al., 1989; Harrison and Burns, 1993). However, the present study does not appear to show such a distinct asymmetry in the tuning of the efferent effect.

Overall, direct comparisons of gain estimates between OAE- and psychophysical-based measures of efferent activation are problematic, even if measuring SFOAEs with elicitors (analogous to a precursor in psychophysical studies) mainly due to constraints in stimulus design. However, once the effects of OAE stimulus constraints such as the use of a broadband elicitor (presented contralaterally to elicit the greatest efferent response), use of an ongoing long-duration probe-tone stimulus (required for pre- and post-elicitor measurement of SFOAEs), and the possibility of two-tone cochlear suppressive effects with ipsilateral elicitor presentation are taken into account, the reduction in SFOAEs with an elicitor (e.g., Guinan et al., 2003) show gain reduction trends that may be in some cases comparable to those found in the present study. The present findings suggest that the 10-dB bandwidth of the efferent-effect filter at 4000 Hz is about 1100 Hz (Q_{10} of 3.5).

5 Conclusions

1. The FDMC method can be used to provide a measure of the frequency tuning of the efferent effect in humans using narrowband ipsilateral precursors.
2. Cochlear gain reduction is greatest for precursors with frequencies close to that of the signal, and least for precursors with frequencies well removed from the signal frequency.
3. The tuning of the efferent effect filter (0.5 to 0.7 octaves above and below the signal frequency) is within the range obtained in humans using OAEs (Lilaonitkul and Guinan, 2009; Zhao and Dhar, 2012).
4. The 10-dB bandwidth of the efferent-effect filter centred at 4000 Hz is about 1100 Hz (Q_{10} of 3.5).

Acknowledgements

The research was supported by an EPSRC grant EP/H022732/1 and an International Project grant awarded by Action on Hearing Loss.

References

ANSI (1996) ANSI S3.6-1996 Specification for Audiometers (American National Standards Institute, New York).

Backus, B. C., & Guinan, Jr. J. J. (2006). Time-course of the human medial olivocochlear reflex. Journal of the Acoustical Society of America, 119, 2889-2904.

Bacon, S. P., & Moore, B. C. J. (1986). Temporal effects in simultaneous pure-tone masking: effects of signal frequency, masker/signal frequency ratio, and masker level. Hearing Research, 23, 257-266.

Bacon, S. P., & Viemeister, N. F. (1985). The temporal course of simultaneous tone-on-tone masking. Journal of the Acoustical Society of America, 78, 1231-1235.

Guinan, J. J. Jr. (2006). Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans. Ear and Hearing, 27, 589-607.

Guinan, J. J. Jr. (2010). Cochlear efferent innervation and function. Current Opinion in Otolaryngology & Head and Neck Surgery, 18, 447-453.

Guinan, J. J. Jr., Backus, B. C., Lilaonitkul,W., & Aharonson, V. (2003). Medial olivocochlear efferent reflex in humans: Otoacoustic emission (OAE) measurement issues and the advantages of stimulus frequency OAEs. Journal of the Association for Research in Otolaryngology, 4, 521-540.

Harrison, W. A., Burns, E. M. (1993). Effects of contralateral acoustic stimulation on spontaneous otoacoustic emissions. Journal of the Acoustical Society of America, 94, 2649-2658.

Huffman, R., & Henson, O. W. (1990). The descending auditory pathway and acousticomotor systems: connections with the inferior colliculus. Brain Research Reviews, 15, 295-323.

James, A. L., Mount, R. J., Harrison, R. V., (2002). Contralateral suppression of DPOAE measured in real time. Clinical Otolaryngology, 27, 106-112.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 49, 467-477.

Liberman, M. C. (1988) Response properties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise. Journal of Neurophysiology, 60, 1779-1798.

Lilaonitkul, W., & Guinan J. J. (2009) Human medial olivocochlear reflex: Effects as functions of contralateral, ipsilateral, and bilateral elicitor bandwidths. Journal of the Association for Research in Otolaryngology, 10, 459-470.

Maison, S., Micheyl, C., Andeol, G., Gallego, S., & Collet, L. (2000). Activation of medial olivocochlear efferent system in humans: influence of stimulus bandwidth. Hearing Research, 140, 111-125.

Mott, J. B., Norton, S. J., Neely, S. T., & Warr, B. (1989). Changes in spontaneous otoacoustic emissions produced by acoustic stimulation of the contralateral ear. Hearing Research, 38, 229-242.

Murugasu, E., & Russell, I. J., (1996). The effect of efferent stimulation on basilar membrane displacement in the basal turn of the guinea pig cochlea. Journal of Neuroscience, 16, 325-332.

Russell, I. J., & Murugasu, E. (1997). Efferent inhibition suppresses basilar membrane responses to near characteristic frequency tones of moderate to high intensities. Journal of the Acoustical Society of America, 102, 1734-1738.

Yasin, I., Drga, V., & Plack, C. J. (2013a). Estimating peripheral gain and compression using fixed-duration masking curves. Journal of the Acoustical Society of America, 133, 4145-4155.

Yasin, I., Drga, V., Plack, C. J. (2013b). Improved psychophysical methods to estimate peripheral gain and compression. In B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, & H. E. Gockel (Eds), *Basic Aspects of Hearing: Physiology and Perception* (pp. 39-46). Heidelberg: Springer-Verlag.

Yasin, I., Drga, V., & Plack, C. J. (2014). Effect of human auditory efferent feedback on cochlear gain and compression. Journal of Neuroscience, 34, 15319-15326.

Zhao, W., & Dhar, S. (2012). Frequency tuning of the contralateral medial olivocochlear reflex in humans. Journal of Neurophysiology, 108, 25-30.