SPEECH INTELLIGIBILITY, SPATIAL UNMASKING, AND REALISM IN REVERBERANT SPATIAL AUDITORY DISPLAYS

Barbara Shinn-Cunningham

Boston University Hearing Research Center,
Departments of Cognitive and Neural Systems and Biomedical Engineering
677 Beacon St., Boston, MA 02215
shinn@cns.bu.edu

ABSTRACT

Many auditory displays strive to include accurate directional spatial cues, but few provide robust cues for source distance. This paper considers how including echoes and reverberation in a spatial auditory display (in order to create salient cues for source distance) impacts other aspects of performance, especially speech intelligibility and spatial unmasking. Preliminary results from masked speech intelligibility studies (together with results from previous experiments investigating sound localization) suggest that including modest amounts of reverberation (such as that present in a typical, everyday room) can provide useful distance information without causing large performance degradations on other tasks.

1. INTRODUCTION

Many researchers have examined how directional sound location cues can be simulated in a spatial auditory display [1-7], but relatively less attention has been given to simulating source distance (although see [8-12]). The most robust cues for source distance (particularly for unfamiliar sources) arise from echoes and reverberation [8-19]. While the ability to judge source distance can be predicted by considering the reverberation and echoes present [20], the actual neural mechanism(s) by which source distance is computed is not known

This paper focuses on how including reverberation and echoes in a spatial auditory simulation will impact other aspects of auditory perception, especially speech perception in the presence of a competing masking source. The remaining sections of this paper describe some basic ideas relevant for understanding how spatial information may influence masked speech intelligibility in reverberant settings. Preliminary results from perceptual studies of masked speech intelligibility in a spatial auditory display are then presented. The final section summarizes these results in light of previous investigations of directional localization accuracy in reverberant rooms.

2. SPATIAL UNMASKING

Spatial auditory cues not only provide a listener with information about sound source location (a useful result unto itself), they can allow a listener to better monitor simultaneous sources when the sources are at different spatial locations [21, 22]. The phrase spatial unmasking refers to the improvements in thresholds for masked source detection and feature discrimination (including improvements in masked speech reception thresholds) that arise when a target sound source and an interfering masker are at different locations in space (relative to when target and masker are at the same location).

In general, spatial unmasking arises due to both pure energetic effects and spatial or binaural processing (e.g., see [22-24]). Energetic effects arise due to the fact that when target and masker are at the same location in space, the target-to-masker energy ratio (TMR) is equal at both ears; however, if the target (or masker) is displaced, the TMR will generally increase at one ear (the better ear) and decrease at the other ear (the worse ear). In addition, even if one takes into account the energetic change in TMR at the better ear, additional spatial unmasking arises (that cannot be explained by changes in the better-ear TMR) when the target and masker give rise to different interaural time or level differences. For detection of low-frequency signals, these spatial effects can be as large as 15 dB [25-28]. Spatial effects can lead to as much as 6 dB of unmasking on speech intelligibility tasks when the masker is a steady-state noise (that is perceptually easy to distinguish from the target) [22-24, 29-31]. For tasks in which the target and masker are difficult to segregate (in cases of so-called informational masking), spatial processing can lead to 15 dB of unmasking [32-36], even for speech signals (where the most important information is at relatively high frequencies between 2-5 kHz where binaural processing advantages are smaller than at lower frequencies).

Because spatial unmasking can provide large improvements on behavioral tasks, many auditory displays are designed to provide accurate directional spatial cues, thus allowing listeners to make use of natural spatial processing mechanisms for monitoring multiple sources.

3. ECHOES AND REVERBERATION

Currently, there is no consensus on how the auditory system computes source distance from reverberant signals. However, it is clear that the relative strength of reverberation (compared to the direct sound energy reaching the listener) changes systematically with source distance [10, 18, 20, 37]. This change in the direct-to-reverberant energy ratio causes concomitant changes in many acoustic properties of the signals reaching the listener, including interaural correlation, temporal modulation, and spectral content. Any or all of these attributes may be used by the auditory system to compute source distance; more research is needed to determine which acoustic attributes due to the reverberation are perceptually relevant.

Although including realistic echoes and reverberation in a display improves perception of source distance, it causes small but measurable degradations in perception of source direction [38, 39]. The fact that echoes and reverberation distort directional hearing is not surprising, because echoes and reverberation distort interaural time differences (ITDs),

interaural level differences (ILDs), and spectral shape, the main cues for source direction [10, 18, 37, 40]. Previous results examining how well subjects localize in rooms (conducted in a moderate-sized classroom with broadband T_{60} = 650 ms) show that directional accuracy is only modestly degraded (mean localization errors are increased by roughly 25%), but distance perception is significantly enhanced (by an order of magnitude) compared to in anechoic space [39, 40].

In addition to influencing spatial perception, echoes and reverberation alter the temporal modulations in the signal reaching a listener. In particular, echoes and reverberation tend to temporally smear out amplitude modulations, particularly at higher modulation frequencies [41-46].

4. SPEECH IN REALISTIC ROOMS

For quasi-steady-state portions of a speech signal, such as vowels, the main acoustic features are conveyed by the relative energy content at each frequency (which is roughly constant over the vowel duration). However, for most other speech sounds, information is conveyed through changes in energy over time and frequency; i.e., much of the information in a speech signal is conveyed by temporal modulations in the energy of envelope of the speech signal at each frequency [47-49].

Because echoes and reverberation can reduce these temporal modulations, echoes and reverberation can degrade speech intelligibility in some acoustic environments. However, for most ordinary (i.e., relatively small rooms), the temporal extent of echoes and reverberation is short compared to the modulations in speech, and only modest perceptual degradations arise, at least at the ear receiving the more intense direct sound (e.g., see [42]). Of course, the severity of the effects of echoes and reverberation on the signals at the ears varies with the location of the source relative to the listener because the direct sound level varies with direct and distance.

These effects are demonstrated in Figure 1, which plots a sample of a speech signal reaching the left ear in anechoic space (in black) superimposed over the signal that would reach the ear in a normal (moderate-sized) classroom (plotted in gray) for a source at a distance of 1 m and azimuth of 90ß to the right (in the horizontal plane containing the ears). These results were generated by measuring the head-related impulse responses (HRIRs) in the classroom using a maximum-length sequence technique, then processing raw speech waveforms through either pseudo-anechoic HRIRs (in which echoes and reverberation were removed through time windowing) or reverberant HRIRs (in which both the direct and reverberant cues were included).

Results show that echoes and reverberation have the largest effect on the total signal at the ear when the source is at 90ß to the right and the left ear signal is considered. In these cases, the direct sound energy is relatively low, leading to large influences of the echoes and reverberation.

5. SPATIAL UNMASKING OF SPEECH IN ROOMS

Echoes and reverberation cause degradations in both directional hearing and speech intelligibility; thus, echoes and reverberation may degrade the benefit of spatial separation of target and masker sources on speech intelligibility.

In order to examine how realistic room echoes and reverberation influence spatial unmasking, a study was conducted under headphones. Target and masker signals

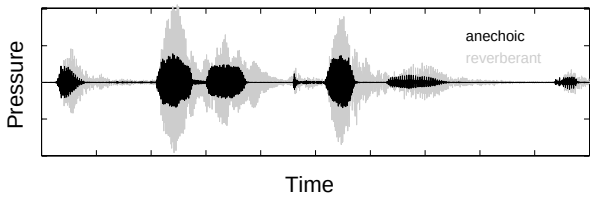

Figure 1. *Sample speech signal reaching the left ear (from a position of 1 m, 90ß) with and without reverberation. The reverberated signal shows less extreme modulation than the anechoic signal*

were simulated at different locations using the pseudo-anechoic and reverberant HRIRs used to process the signal shown in Figure 1. The masker was a steady-state noise, which was always simulated at a position directly in front of the listener at a distance of 15 cm. The target signals were nonsense sentences simulated at one of the three distances (0.15, 1, or 2 m) and two directions (0ß and 90ß) for which HRIRs were measured, leading to six different target/masker spatial configurations.

For each spatial configuration, subjects were tested while listening binaurally, with only the left ear, and with only the right ear, to allow direct analysis of the advantages of binaural processing. Two different room conditions were tested for each spatial configuration and ear condition, one simulating anechoic space and one simulating reverberant conditions.

For each condition, speech reception thresholds were measured adaptively by varying the target level until 50% of the sentence key words were understood. Each threshold was measured four times to estimate final thresholds. Four normal-hearing subjects completed each test.

Figure 2 plots the raw thresholds of the direct-sound portion of the target (relative to the direct-sound level of the masker signal) at the 50% speech-reception threshold. Each panel gives results for a different individual listener; within each panel, results are shown as a function of source distance for the different room conditions and target directions.

The plots in Figure 2 show how threshold TMR *at the listener s better ear* changes with spatial configuration of T and M, but not how the level emitted by T would have to change to achieve threshold performance. Put another way, because results are plotted in terms of the TMR at the better ear at threshold, any changes in the TMR that would arise due to changes in spatial configuration are hidden. For instance, there are very large decreases in the target level reaching the listener as the target moves from very near the head to a distance of 1 m; however, this overall energy change is removed given how results are plotted. Similarly, the TMR at the right (better) ear increases when T is moved to 90ß relative to when T is at 0..azimuth. However, this energetic change in the better-ear TMR is removed in Figure 2. Thus, the plots in Figure 2 generally underestimate the magnitude of spatial unmasking effects that would obtain in the real world because, in the plots, energetic effects are (at least crudely) normalized out. This method for plotting the data was chosen because it emphasizes differences in performance that arise beyond obvious energetic effects.

Additionally, the method used for normalizing the results in Figure 2 ignores the reverberant energy from T and M when computing the TMR at the better ear. Thus, to the extent that there are differences in TMR in the anechoic and reverberant conditions, the plot shows the total effect of adding walls to the listening environment (relative to the condition when there are no reflective surfaces).

Overall, the pattern of results is very similar across the four subjects. Comparing the better-ear (right ear; dotted line) and binaural (solid line) results, the data show that directional separation of target and masker leads to binaural processing advantages of 3-5 dB in both anechoic and reverberant simulations. Thus, the interaural decorrelation of the target and masker signals does not cause any significant decrease in the effectiveness of binaural processing.

When target and masker are in the same direction in anechoic space, there is no significant or consistent difference across performance achieved with the left ear alone, right ear alone, or when listening binaurally. However, in the reverberant simulations, there is a distinct binaural processing advantage when the target is at a different distance than the masker.

When considering the conditions in which T was to the right, comparisons between the left (worse) and right (better) ear results show very large differences in monaural performance. Given the way data are normalized, this difference primarily reflects the large interaural level differences (ILDs) in T that occur when T is near and to the side of the listener [50-52]; this large ILD decreases with distance, leading to corresponding decrements in the difference in the left and right ear monaural thresholds with distance. Comparing anechoic and reverberant results, Figure 2 shows that the addition of echoes and reverberation tends to decrease the differences in left- and right-ear monaural thresholds, especially at the farthest distance (where the reverberation has the largest impact). To the extent that reflected target energy is helpful rather than detrimental to understanding the target, this effect is easily explained. Specifically, the reflected target energy is (at least to a first-order approximation) roughly equal at the two ears for all conditions, whereas the direct sound ILD in the target is quite large for T near and to the right of the listener. The echoes and reverberation thus tend to have a large impact on monaural intelligibility for the acoustically-worse ear, where there is very little T energy reaching the listener in anechoic space. Overall, then, the echoes and reverberation tend to reduce the better ear advantage by disproportionately improving performance for the acoustically worse ear.

These results indicate that spatial unmasking of speech is not only effective in normal reverberant rooms, but that echoes and reverberation can actually lead to improvements in binaural processing (e.g., when T and M are both in front of the listener, but T is relatively far and M is near). This improvement is probably due to decorrelation of the target signal in the presence of the masker (which is sufficiently close to the listener that the effect of reverberation and echoes on the interaural cues is quite small, leading to an essentially diotic masking signal). Further, reverberant energy can lead to improvements in monaural performance by boosting the effective TMR at the acoustically worse ear. Thus, moderate levels of reverberation (such as occur in an ordinary classroom) either lead to improvements or no noticeable change in both binaural and monaural conditions.

6. CONCLUSIONS

Results of the spatial unmasking study suggest that in a moderate-sized classroom, binaural processing is as effective as in anechoic space when target and masker are separated in direction. In addition, differences in distance in a reverberant room can also lead to binaural processing advantages that are as large as the advantages due to directional separation. These spatial unmasking results simulated the same reverberant space in which previous real-world localization studies were performed. In this space, reverberation and echoes allowed subjects to judge source distance with good accuracy.

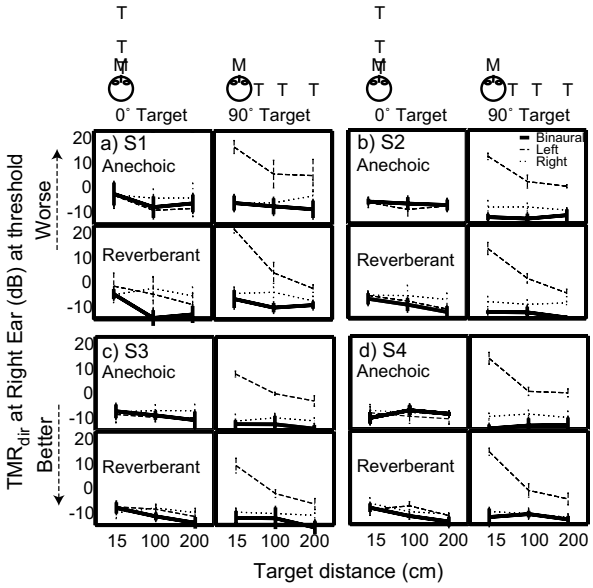

Figure 2. *Mean TMR in dB RMS at 50% words correct threshold. Error bars show within-subject std. dev. Direct-sound TMR is fixed at the right ear to illustrate the better-ear advantage; this analysis ignores any positive contributions of T reverberation. Spatial configuration of target and masker indicated by cartoons at top of figure.*

These results demonstrate that including realistic reverberation in spatial auditory displays improve distance perception and may even increase spatial unmasking at a cost of modest (negligible) degradations in directional localization accuracy. Because realistic echoes and reverberation also lead to very large improvements in the subjective realism of spatial auditory displays, most spatial auditory displays should include echoes and reverberation. The only compelling reasons for not including realistic room acoustics in spatial auditory displays may arise from practical and technical constraints on the simulation that can be achieved and the amount of processing power that can be afforded in building a real-time display.

Further work is necessary in order to identify what aspects of echoes and reverberation are critical for supplying distance information and realism in spatial auditory displays. Such knowledge may allow future displays to incorporate simple reverberation models that yield the benefits of realistic reverberation with reduced computational complexity. Similarly, further work investigating the effects of reverberation on speech intelligibility and spatial unmasking will allow designers to make informed choices about how much reverberation to include, and at what cost.

7. ACKNOWLEDGEMENTS

This work was supported in part AFOSR Grant No. F49620-01-1-0005 and the Alfred P. Sloan Foundation. Scarlet Constant and Norbert Kopco assisted with data collection. For additional information and related papers, see http://www.cns.bu.edu/~shinn/.

8. REFERENCES

1. Wenzel, E.M., *Localization in virtual acoustic displays.* Presence, 1992. **1**(1): p. 80-107.
2. Wightman, F.L. and D.J. Kistler, *Headphone simulation of free-field listening. II. Psychophysical validation.* Journal of the Acoustical Society of America, 1989. **85**: p. 868-878.
3. Wightman, F.L., D. Kistler, and P. Zahorik, *Issues and non-issues in the production of high-resolution auditory virtual environments*, in *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality*, M. Smith, et al., Editors. 2001, Lawrence Erlbaum: New Jersey. p. 594-598.
4. Langendijk, E.H.A. and A.W. Bronkhorst, *Fidelity of three-dimensional-sound reproduction using a virtual auditory display.* Journal of the Acoustical Society of America, 2000. **107**(1): p. 528-537.
5. Bronkhorst, A.W., *Localization of real and virtual sound sources.* Journal of the Acoustical Society of America, 1995. **98**(5): p. 2542-2553.
6. Carlile, S. and J. Leung, *Rendering sound sources in high fidelity virtual auditory space: Some spatial sampling and psychophysical factors*, in *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality*, M. Smith, et al., Editors. 2001, Lawrence Erlbaum: New Jersey. p. 599-603.
7. Carlile, S., C. Jin, and V. Harvey, *The generation and validation of high fidelity virtual auditory space.* Proceedings of the 20th International Conference of the IEEE Engineering in Medicine and Biology Society, 1998. **20**(3/6): p. 1090-1095.
8. Begault, D.R., et al., *Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source.* Journal of the Audio Engineering Society, 2001. **49**(10): p. 904-916.
9. Zahorik, P., *Loudness constancy with varying sound source distance.* Nature Neuroscience, 2001. **4**(1): p. 78-83.
10. Shinn-Cunningham, B.G. *Distance cues for virtual auditory space*. in *Proceedings of the IEEE-PCM 2000*. 2000. Sydney, Australia.
11. Shinn-Cunningham, B.G., *Creating three dimensions in virtual auditory displays*, in *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality*, M. Smith, et al., Editors. 2001, Lawrence Erlbaum: New Jersey. p. 604-608.
12. Brungart, D.S., *Near-field virtual audio displays.* Presence, 2002. **11**(1): p. 93-106.
13. Mershon, D.H. and L.E. King, *Intensity and reverberation as factors in auditory perception of egocentric distance.* Perception and Psychophysics, 1975. **18**: p. 409-415.
14. Mershon, D.H. and J.N. Bowers, *Absolute and relative cues for the auditory perception of egocentric distance.* Perception, 1979. **8**: p. 311-322.
15. Mershon, D.H., et al., *Effects of room reflectance and background noise on perceived auditory distance.* Perception, 1989. **18**: p. 403-416.
16. Begault, D.R., *Perceptual effects of synthetic reverberation on three-dimensional audio systems.* Journal of the Audio Engineering Society, 1992. **40**(11): p. 895-904.
17. Begault, D.R., B.U. McClain, and M.R. Anderson. *Early reflection thresholds for virtual sound sources*. in *2001 International Workshop on Spatial Media*. 2001. Aizu-Wakamatsu, Japan.
18. Shinn-Cunningham, B.G. *Creating three dimensions in virtual auditory displays*. in *Proceedings of HCI International 2001*. 2001. New Orleans.
19. Martens, W.L. and A. Yoshida. *Psychoacoustically-based control of auditory range: Display of virtual sound sources in the listener's personal space*. in *International Conference on Information Society in the 21st Century: Emerging Technologies and New Challenges (ISO2000)*. 2000. Aizu-Wakamatsu, Japan.
20. Bronkhorst, A.W. and T. Houtgast, *Auditory distance perception in rooms.* Nature, 1999. **397**(11 February): p. 517-520.
21. Bronkhorst, A.W., *The cocktail party effect: Research and applications.* Journal of the Acoustical Society of America, 1999. **105**(2): p. 1150.
22. Bronkhorst, A.W., *The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions.* Acustica, 2000. **86**: p. 117-128.
23. Zurek, P.M., *Binaural advantages and directional effects in speech intelligibility*, in *Acoustical Factors Affecting Hearing Aid Performance*, G. Studebaker and I. Hochberg, Editors. 1993, College-Hill Press: Boston, MA.
24. Shinn-Cunningham, B.G., et al., *Spatial unmasking of nearby speech sources in a simulated anechoic environment.* Journal of the Acoustical Society of America, 2001. **110**(2): p. 1118-1129.
25. van de Par, S. and A. Kohlrausch, *Dependence of binaural masking level differences on center frequency, masker bandwidth, and interaural parameters.* Journal of the Acoustical Society of America, 1999. **106**(4): p. 1940-1947.
26. Gilkey, R.H., D.E. Robinson, and T.E. Hanna, *Effects of masker waveform and signal-to-masker phase relation on diotic and dichotic masking by reproducible noise.* Journal of the Acoustical Society of America, 1985. **78**(4): p. 1207-1219.
27. Gilkey, R.H. and M.D. Good, *Effects of frequency on free-field masking.* Human Factors, 1995. **37**(4): p. 835-843.
28. Good, M.D., R.H. Gilkey, and J.M. Ball, *The relation between detection in noise and localization in noise in the free field*, in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. Anderson, Editors. 1997, Erlbaum: New York. p. 349-376.
29. Drullman, R. and A.W. Bronkhorst, *Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation.* Journal of the Acoustical Society of America, 2000. **107**: p. 2224-2235.
30. Bronkhorst, A.W. and R. Plomp, *A clinical test for the assessment of binaural speech perception in noise.* Audiology, 1990. **29**: p. 275-285.
31. Bronkhorst, A.W. and R. Plomp, *The effect of head-induced interaural time and level differences on speech intelligibility in noise.* Journal of the Acoustical Society of America, 1988. **83**: p. 1508-1516.
32. Freyman, R.L., U. Balakrishnan, and K. Helfer. *Release from informational masking in speech recognition*. in *MidWinter Meeting of the Association for Research in Otolaryngology*. 2000. St. Petersberg Beach, FL.
33. Freyman, R.L., et al., *The role of perceived spatial separation in the unmasking of speech.* Journal of the Acoustical Society of America, 1999. **106**(6): p. 3578-3588.

34. Arbogast, T.L. and J. Kidd, Gerald, *Evidence for spatial tuning in informational masking using the probe-signal method.* Journal of the Acoustical Society of America, 1999: p. submitted.
35. Brungart, D.S., *Informational and energetic masking effects in the perception of two simultaneous talkers'.* Journal of the Acoustical Society of America, 2001. **109**(3): p. 1101-1109.
36. Kidd, G., et al., *Reducing informational masking by sound segregation.* Journal of the Acoustical Society of America, 1994. **95**(6): p. 3475-3480.
37. Shinn-Cunningham, B.G., J.G. Desloge, and N. Kopco. *Empirical and modeled acoustic transfer functions in a simple room: Effects of distance and direction*. in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2001. New Pfalz, New York.
38. Santarelli, S., *Auditory Localization of Nearby Sources in Anechoic and Reverberant Environments*, in *Cognitive and Neural Systems*. 2000, Boston University: Boston, MA.
39. Shinn-Cunningham, B.G. *Learning reverberation: Implications for spatial auditory displays*. in *International Conference on Auditory Displays*. 2000. Atlanta, GA.
40. Shinn-Cunningham, B.G. *Localizing sound in rooms*. in *ACM/SIGGRAPH and Eurographics Campfire: Acoustic Rendering for Virtual Environments*. 2001. Snowbird, Utah.
41. Houtgast, T. and H.J.M. Steeneken, *A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria.* Journal of the Acoustical Society of America, 1985. **77**(3): p. 1069-1077.
42. Houtgast, T., H.J.M. Steeneken, and R. Plomp, *Predicting speech intelligibility in rooms from the modulation transfer function I. General room acoustics.* Acustica, 1980. **46**: p. 60-72.
43. Nomura, H., H. Miyata, and T. Houtgast, *Speech-intelligibility and subjective MTF under diotic and dichotic listening conditions in reverberant sound fields.* Acustica, 1991. **73**(4): p. 200-207.
44. Plomp, R., H.J.M. Steeneken, and T. Houtgast, *Predicting speech intelligibility in rooms from the modulation transfer function II. Mirror image computer model applied to rectangular rooms.* Acustica, 1980. **46**: p. 73-81.
45. Payton, K.L., R.M. Uchanski, and L.D. Braida, *Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing.* Journal of the Acoustical Society of America, 1994. **95**(3): p. 1581-1592.
46. Payton, K.L. and L.D. Braida, *A method to determine the speech transmission index from speech waveforms.* Journal of the Acoustical Society of America, 1999. **106**(6): p. 3637-3648.
47. Greenberg, S. and B.E.D. Kingsbury. *The modulation spectrogram: In pursuit of an invariant representation of speech*. in *ICASSP-97*. 1997. Munich.
48. Greenberg, S. and T. Arai. *The relation between speech intelligibility and the complex modulation spectrum*. in *7th International Conference on Speech Communication and Technology*. 2001.
49. Chi, T., et al., *Spectro-temporal modulation transfer functions and speech intelligibility.* Journal of the Acoustical Society of America, 1999. **106**(5): p. 2719-2732.
50. Duda, R.O. and W.L. Martens, *Range dependence of the response of a spherical head model.* Journal of the Acoustical Society of America, 1998. **104**(5): p. 3048-3058.
51. Brungart, D.S. and W.M. Rabinowitz, *Auditory localization of nearby sources I: Head-related transfer functions.* Journal of the Acoustical Society of America, 1999. **106**(3): p. 1465-1479.
52. Shinn-Cunningham, B.G., S. Santarelli, and N. Kopco, *Tori of confusion: Binaural localization cues for sources within reach of a listener.* Journal of the Acoustical Society of America, 2000. **107**(3): p. 1627-1636.