THE EFFECT OF AUDITORY SPATIAL LAYOUT IN A DIVIDED ATTENTION TASK

Virginia Best, Antje Ihlefeld and Barbara Shinn-Cunningham

Boston University Hearing Research Center
Department of Cognitive & Neural Systems
677 Beacon St, Boston, MA, 02215, USA
<ginbest,ihlefeld,shinn>@bu.edu

ABSTRACT

The effect of spatial separation on the ability of listeners to report keywords from two simultaneous talkers was examined. The talkers were presented with equal intensity at a clearly audible level, and were designed to have little spectral overlap in order to reduce energetic interference. The two talkers were presented in a virtual auditory environment with various angular separations around references of -45º, 0º, or 45º azimuth. In Experiment 1, the virtual space was created using head-related transfer functions (HRTFs) which contained natural energy variations as a function of location. In Experiment 2, these energy variations were removed and the virtual space was created using only interaural time differences (ITDs). Overall, performance did not vary dramatically but depended on spatial separation, reference direction, and type of simulation. Around the 0º reference azimuth, performance in the HRTF condition tended to first increase and then decrease with increasing separation. This effect was greatly reduced in the ITD condition and thus appears to be related primarily to energy variations at the two ears. For sources around the ± 45º reference azimuths, there was an advantage to separating the two sources in both HRTF and ITD conditions, suggesting that perceived spatial separation is advantageous in a divided attention task, at least for lateral sources.

1. INTRODUCTION

Many studies have examined the role of auditory spatial layout in selective attention situations, where a listener must extract the content of one source (a 'target') in the presence of competing sources ('maskers') [see 1 for review]. In selective attention tasks, separation of the target from the maskers can improve performance. When the masker reduces the audibility of components of the target ('energetic masking'), there are two ways in which spatial separation offers an advantage. First, the relative energy of the target and masker at the ears changes with target and masker location, increasing the target audibility in each frequency band at one of the ears. Second, binaural processing allows listeners to detect the presence of target energy in a particular band if the target and masker contain different interaural time and/or level differences [1, 2]. When competing sources do not have significant frequency overlap and reduced audibility is not the primary source of interference, a masker with similar spectro-temporal characteristics can still interfere with the perception of the target. Such so-called 'informational masking' is also greatly reduced by spatial separation of the target and masker. In these conditions it is thought that the differences in perceived location reduce the confusability of the two sources, and allow listeners to direct their attention selectively to the target [e.g. 3-5].

In many situations, it is necessary for a listener to follow more than one sound source at a time. For example, in an auditory display with two competing talkers, it may be important that information be extracted from each talker. Previous studies of divided attention between speech sources have focused on monaural or diotic presentation [e.g. 6] and have not considered spatial factors in detail. In a divided attention task, it is not clear what the effect of spatial separation of the two targets might be. It is reasonable to expect that spatial separation would be advantageous in that it would enhance the audibility of the two sources, as well as reducing confusion between them, as described above. However, if one considers the issue of directing spatial attention to relevant sound sources, spatial separation could in fact be detrimental in a divided attention task. If spatial attention acts as a 'spotlight', this spotlight may have to be broadened to simultaneously encompass two targets. On the other hand, if the spotlight of attention is relatively narrow, a divided attention task may involve rapid switching between the two target locations. In either case, it may be advantageous to have the two sources of interest within a restricted region, i.e. it may be more difficult to divide (or switch) attention between spatially separated targets (see Figure 1).

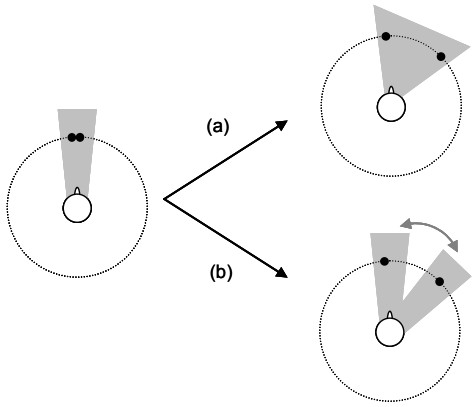

Figure 1. *If auditory spatial attention acts as a 'spotlight' to enhance the perception of a relevant source and exclude others, it may be advantageous to have any sources of interest within a restricted region. (a) If listeners divide attention by enlarging the spotlight, there is presumably a limit on how much it can be enlarged without sacrificing accuracy. (b) If a switching strategy is adopted in divided attention tasks, it may be increasingly difficult to do so as sources are separated.*

This study was designed to examine systematically the effect of spatial configuration on the ability of listeners to report keywords from two simultaneous talkers. In order to emphasise the influence of spatial attention on performance, an effort was made to minimise the contribution of energetic masking. Energetic interference between the two sources was minimised by (a) presenting the two talkers with equal intensity at a clearly audible level, and (b) processing the two speech signals to minimise their spectral overlap (as in [7]; see section 2.2 for details).

2. METHODS

2.1. Subjects

Eight paid subjects (ages 20 – 30) participated in the study. Four subjects had previous experience in psychophysical studies of a similar nature. All subjects participated in Experiment 1, and six of the subjects went on to participate in Experiment 2.

2.2. Stimuli

The speech materials consisted of spoken sentences that were taken from the publicly available Coordinate Response Measure speech corpus [8]. These sentences all contain seven words, three of which are keywords that vary from utterance to utterance. The form of the sentences is "Ready *call-sign* go to *colour number* now", where the italicised words indicate keywords. In the corpus there are eight possible call-signs ("arrow", "baron", "charlie", "eagle", "hopper", "laker", "ringo", "tiger"), four possible colours ("blue", "green", "red", "white"), and eight possible numbers (1-8). All combinations of these words produce 256 phrases, which are each spoken by eight talkers (four male, four female), giving a total of 2048 sentences. The sentences are time-aligned such that the word "ready" always starts at the same time, but some variations in overall rhythm occur between different sentences, so that the keywords in different utterances are not exactly aligned.

For each trial, two sentences spoken by the same talker were chosen randomly from the corpus with the restriction that all keywords differed in the two sentences. In order to reduce the energetic interference between the two sentences, they were processed to produce intelligible speech-like signals that had little spectral overlap [7, 9]. The signals were band-pass filtered into 8 non-overlapping frequency bands of 1/3 octave width, with centre frequencies spaced evenly on a logarithmic scale from 175 to 4400 Hz. Four bands were randomly chosen for the first sentence (two from the four lower bands and two from the four higher bands). The Hilbert-envelope of each band was then used to modulate a sinusoidal carrier at the centre frequency of that band, and the sentence was reconstructed by summing the four modulated sinusoids. For the second sentence, the remaining four frequency bands were chosen and the same procedure was followed. The two reconstructed sentences were RMS-normalised to result in a relative level of 0 dB (see Figure 2 for example spectra).

The stimuli were processed to create binaural signals containing realistic spatial cues, and presented over headphones. In Experiment 1, a full set of spatial cues was used in the simulation. Binaural stimuli were created by convolving the speech signal with the appropriate anechoic left and right head-related transfer functions (HRTFs) measured on a KEMAR manikin at a distance of 1 metre [10]. The two binaural stimuli were then added to simulate the two speech sources at their desired locations in external space. In Experiment 2, energy differences between the ears that were present in the HRTF simulation were removed in order to eliminate location-dependent variations in the relative levels of the two sentences. Thus only one spatial cue (the interaural time difference, ITD) was used in these simulations. Appropriate ITDs, extracted from the HRTFs by finding the time-delay of the peak in the broadband interaural cross-correlation function, were used to delay the left and right ear signals.

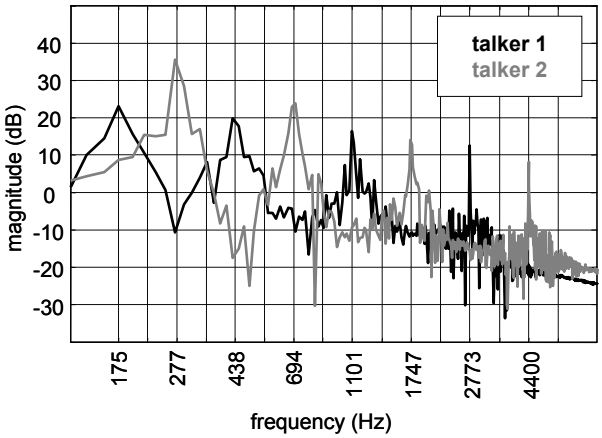

Figure 2. *Frequency spectrum of two example sentences after processing. The two signals were processed to minimise their spectral overlap. Sentences were divided into eight 1/3 octave bands with centre frequencies between 175 and 4400 Hz. Four different bands were chosen for each sentence and their envelopes used to modulate sinusoids at the centre frequency of each band. Intelligible speech signals were reconstructed by summing the four modulated sinusoids.*

Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. Digital stimuli were sent to Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over insert headphones (Etymotic Research ER-2). Subjects were seated in a sound treated booth in front of the PC terminal displaying a graphical user interface (GUI). Following each presentation, subjects indicated their responses by clicking on the GUI, allowing the PC to store their responses.

2.3. Training

Before commencing each of the experiments, subjects participated in a short series of training runs designed to familiarise them with the stimuli and task. In a training test, subjects were presented with stimuli containing a single sentence in quiet, and were required to indicate the colour/number pair they perceived. After each trial, correct-answer feedback was provided by a written message on the screen. A training run consisted of 130 trials. Subjects completed as many runs as required to bring their proportion of correct responses to at least 95%. All subjects reached this level within three training runs.

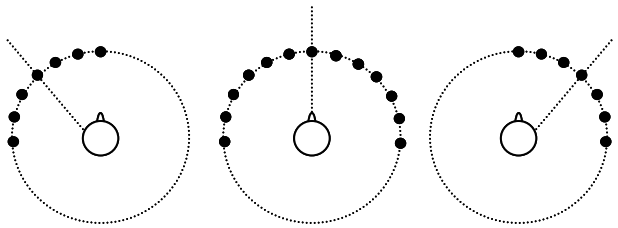

Figure 3. *Spatial configurations tested. The two targets were separated symmetrically about three reference azimuths: -45°, 0°, and 45°. For the frontal reference location, separations of 0°, 30°, 60°, 90°, 120°, 150° and 180° were tested. For the lateral locations, separations of 0°, 30°, 60° and 90° were tested.*

2.4. Experimental Design

In the experimental sessions, two simultaneous sentences were presented and subjects were required to respond with *two* colour/number pairs (in either order). No feedback was provided. A response was considered correct only if both colour/number pairs were reported correctly. Note that chance performance, achieved by randomly guessing the two colour/number pairs, is only 0.3%.

Stimulus locations were all on the horizontal plane passing through the ears (0° elevation) and are described in terms of their angle from the midline (azimuth). Performance was measured with sources separated symmetrically about three reference azimuths (-45°, 0°, and 45°) as illustrated in Figure 3. For the frontal reference location, the two sentences were separated by 0°, 30°, 60°, 90°, 120°, 150° or 180°. For the lateral locations, the two sentences were separated by 0°, 30°, 60° or 90°. Thus a total of 15 unique configurations were examined.

All 15 configurations were presented five times in a random order in each run, for a total of 75 trials per run. Each subject completed 10 such runs for each experiment, and thus gave a total of 50 responses for each configuration. The 20 runs (10 each for Experiments 1 and 2) were carried out over four to five sessions. This meant that subjects did no more than one hour of testing per day and were not fatigued.

3. EXPERIMENT 1

3.1. Mean Percent Correct

Individual subjects differed in their absolute level of performance, but overall trends were similar. Mean percent correct scores across subjects (and standard errors) are shown in Figure 4. For the lateral configurations, left-side and right-side results have been collapsed. Spatial configuration had a modest effect on performance; for a given subject performance did not vary by more than 30 percentage points across all configurations. However, there were consistent patterns in the data as a function of the spatial separation. For the frontal reference location, performance tended to first increase and then decrease with increasing source separation, peaking at 90° - 120° separation. With both talkers to one side, a similar trend was observed but with performance peaking at 60° separation.

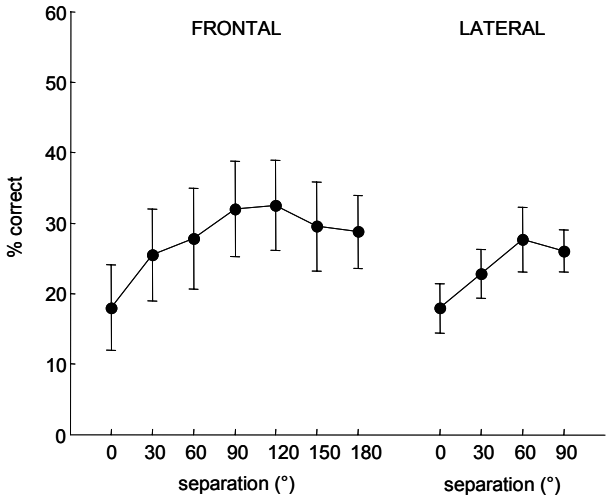

Figure 4. *Mean percent correct scores for the different spatial configurations in Experiment 1. Results are pooled across the eight subjects and error bars indicate standard error of the mean. For the lateral configuration, data have been collapsed across left and right sides.*

3.2. Normalised Results

In order to factor out overall differences in subject performance and concentrate on the effect of spatial separation, percent correct scores for each subject at each reference location were normalised by subtracting the percent correct in the co-located (separation 0°) configuration for that reference location. The resulting normalised values summarise how performance changed with source separation. Figure 5 shows the normalised data pooled across the eight subjects (means and standard errors). The trends described for the raw data are reinforced: increasing the spatial separation tended to cause an increase and then a decrease in performance.

3.3. Acoustic Analysis

Although the two targets were nominally presented with equal intensity, variations in the HRTF for different spatial locations result in variations in the level of each target at each ear. This is especially evident for a target placed to the side, where the acoustic shadow cast by the head greatly attenuates the level received at the far ear, particularly at high frequencies (above about 2 kHz). Indeed for a given spatial configuration, each of the two sources would have a different ear in which its level (relative to the other source) was greater. Moreover, the magnitude of this better ear 'level ratio' would vary as a function of the spatial configuration.

An acoustic analysis was performed to examine whether such level variations might help to explain the trends seen in the behavioural data. For each spatial configuration, fifty speech pairs were generated and the level ratio (LR) in the better ear for each source was calculated using the RMS level of each source after HRTF filtering. The changes in better ear LR as a

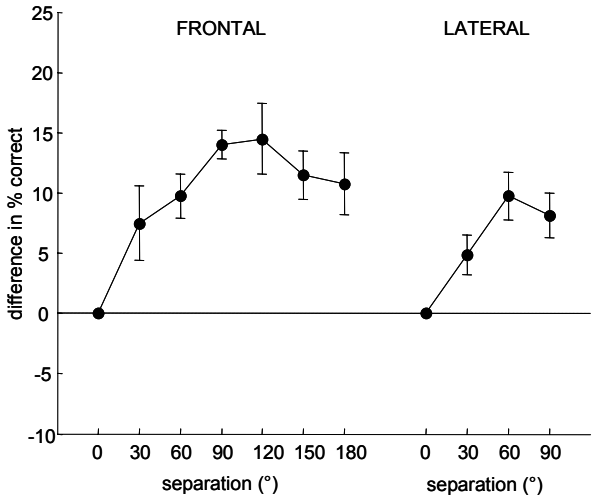

Figure 5. *Mean normalised percent correct scores for Experiment 1. Normalisation was carried out for each individual and each reference location by subtracting the score for the co-located configuration. Results are pooled across the eight subjects and error bars indicate standard error of the mean. For the lateral configuration, data have been collapsed across left and right sides.*

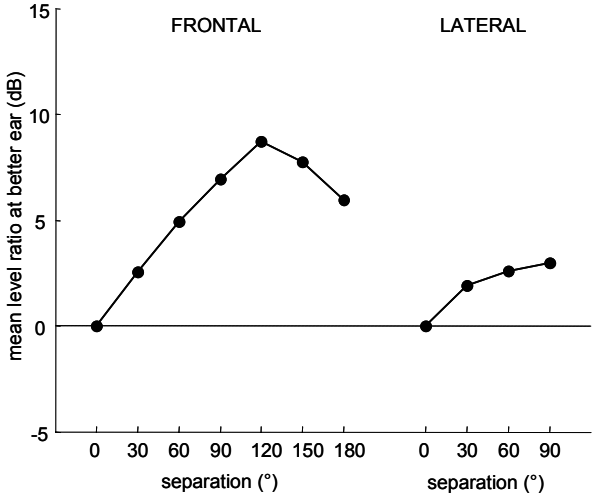

Figure 6. *Level ratios for the different spatial configurations. Level ratios describe the level of a source in its 'better ear' relative to the level of the other source. These ratios were calculated for 50 example stimuli and the means across the two sources (at their respective better ears) are shown.*

function of spatial separation (averaged across the two sources and their respective better ears) are shown in Figure 6. Note that by definition, the LR for a co-located pair is 0 dB. For the lateral stimulus configurations, the LR increases with increasing separation. For the frontal configurations, the LR increases as separation grows to 120°, but then decreases with further separation.

This analysis suggests that the relative levels of the two talkers at each ear in each configuration can partially account for the behavioural results. In general, performance was positively correlated with the mean LRs across the two better ears. Experiment 2 was conducted to further test this explanation.

4. EXPERIMENT 2

Experiment 2 was designed to eliminate energy effects in order to confirm their role in the results of Experiment 1 and to determine whether there is any residual influence of perceived spatial separation of the two sources in a divided attention task. By using only ITDs in the spatial simulation, the level variations induced by the HRTF processing in Experiment 1 were removed (in essence, the LRs for these stimuli are fixed at 0 dB). Any remaining effects of spatial configuration are presumed to be due to the perceived lateral positions of the two sources.

4.1. Mean Percent Correct

The mean percent correct scores (and standard errors) for the different configurations in Experiment 2 are shown in Figure 7 (black lines). Mean results from Experiment 1 for the six subjects who completed both experiments are also plotted for comparison (grey lines). For the frontal reference location, the curve is flatter than in Experiment 1, primarily due to an improvement in performance for co-located sources and sources with small spatial separations. With both talkers to one side, the trends are similar to those seen in Experiment 1 although performance does not decrease for the largest separation. Interestingly, overall performance is better (by approximately 5 percentage points) in Experiment 2 than for the same subjects in Experiment 1. It is important to keep in mind, however, that Experiment 2 was conducted after the completion of Experiment 1, and hence subjects were more experienced. Thus a direct comparison of the percent correct scores is problematic, and the normalised results (next section) are perhaps more relevant.

4.2. Normalised Results

Percent correct scores for each subject at each reference location were normalised by subtracting the score in the co-located (separation 0°) configuration. In Figure 8, the normalised data pooled across the six subjects in Experiment 2 are shown (black lines, means and standard errors). The data for this subset of subjects in Experiment 1 are also plotted (grey lines). Unlike in Experiment 1, there was little consistent change in performance with increasing separation for the 0° reference azimuth. However, spatial separation caused an increase in performance for the lateral reference azimuths that was similar to that seen in Experiment 1. Note however that the improvement is gradual and does not peak at 60° separation.

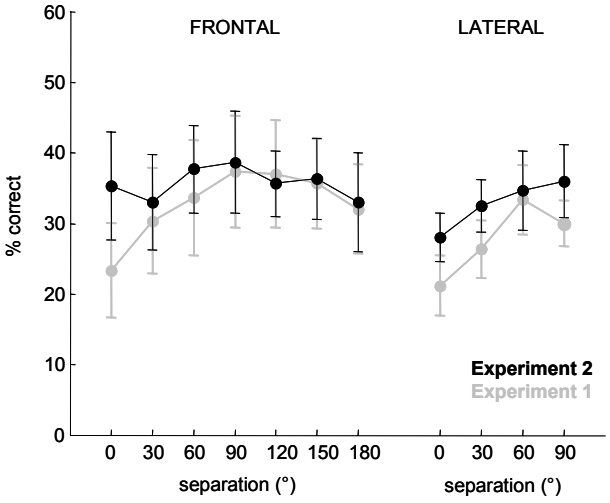

Figure 7. *Mean percent correct scores for the different spatial configurations in Experiment 2 (black lines). Results are pooled across the six subjects and error bars indicate standard error of the mean. For the lateral configuration, data have been collapsed across left and right sides. Mean results for the same six subjects in Experiment 1 are also plotted for comparison (grey lines).*

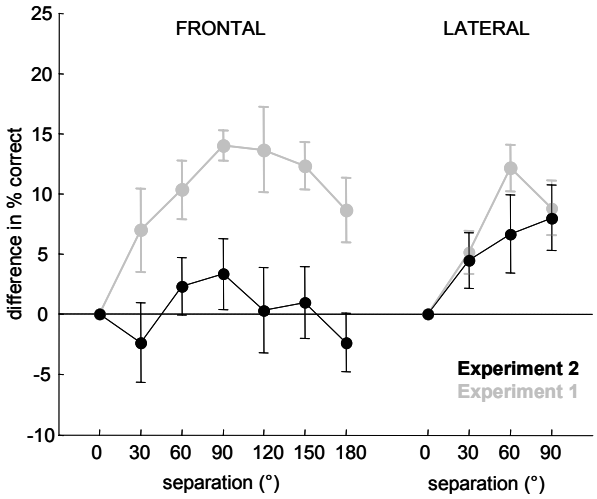

Figure 8. *Mean normalised percent correct scores for Experiment 2 (black lines). Normalisation was carried out for each individual and each reference location by subtracting the score for the co-located configuration. Results are pooled across the six subjects and error bars indicate standard error of the mean. For the lateral configuration, data have been collapsed across left and right sides. Mean results for the same six subjects in Experiment 1 are also plotted for comparison (grey lines).*

5. DISCUSSION AND CONCLUSIONS

This experiment examined the ability of listeners to track two simultaneous speech sources with minimal spectral overlap. It was found that performance varied as a function of the spatial configuration of the two sources.

When sources were separated symmetrically about the frontal midline, performance depended on the type of spatial cues used to simulate the stimulus positions. When full HRTFs were used, performance improved with moderate separations (best performance for separations in the range 90° - 120°) but then decreased with further separation. This trend was much reduced in Experiment 2, when spatial locations were simulated using ITDs only. This suggests that variations in the relative level of the two sources at the ears modulated the difficulty of the task and, ultimately, the accuracy of responses. Indeed, in the best spatial configuration (120° separation), the mean level ratio was 9 dB, meaning that each target source was 9 dB more intense in its better ear than the competing source. This result suggests that listeners use the information at the two ears *independently* when tracking two sources in different hemispheres, an idea that warrants further investigation.

A different story emerges from the results for lateral configurations, where there was an overall benefit from separating the two targets. This spatial benefit cannot be attributed to energy levels, as it was present in the absence of LR changes in Experiment 2. Indeed in Experiment 1, variations in the LRs were relatively small for these lateral configurations (see Figure 6). The persistent separation advantage in the ITD condition suggests that there is an advantage to perceiving the two sources at different lateral positions, at least for lateral sources and the moderate range of separations used in these experiments.

Ultimately, the goal of this work is to understand how spatial attention operates in a divided attention task. The data suggest that, overall, a moderate amount of spatial separation is helpful. This is consistent with the observation that differences in perceived location of competing sources can aid in their segregation and individual intelligibility [3, 4]. However, the results of Experiment 1 suggest that very large separations may be harmful to performance under some circumstances. Although this increasing and then decreasing trend was not evident in the mean data for Experiment 2, it was still prominent in some individuals. This may reflect increased difficulty in following sources that fall outside the optimum range of a putative 'spotlight' of spatial attention. Finally, out of the configurations tested in the present set of experiments, the optimal configuration for dividing attention between a pair of simultaneous sources was with one on each side of the midline in the lateral angle range of 45° - 60°.

6. ACKNOWLEDGMENTS

This work was funded by the Office of Naval Research Grant N00014-04-1-0131.

7. REFERENCES

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica*, vol. 86, pp. 117-128, 2000.

[2] P. M. Zurek, "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Studebaker and I. Hochberg, Eds. Boston: Allyn and Bacon, 1993, pp. 255-276.

[3] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech," *J. Acoust. Soc. Am.*, vol. 106, pp. 3578-3588, 1999.

[4] R. L. Freyman, U. Balakrishnan, and K. S. Helfer, "Spatial release from informational masking in speech recognition," *J. Acoust. Soc. Am.*, vol. 109, pp. 2112-2122, 2001.

[5] T. L. Arbogast and G. Kidd, "Evidence for spatial tuning in informational masking using the probe-signal method," *J. Acoust. Soc. Am.*, vol. 108, pp. 1803-1810, 2000.

[6] D. E. Broadbent, "The role of auditory localization in attention and memory span," *J. Exp. Psychol.*, vol. 47, pp. 191-196, 1954.

[7] T. L. Arbogast, C. R. Mason, and G. Kidd, "The effect of spatial separation on informational and energetic masking of speech," *J. Acoust. Soc. Am.*, vol. 112, pp. 2086-2098, 2002.

[8] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, vol. 107, pp. 1065-1066, 2000.

[9] D. S. Brungart, B. D. Simpson, C. J. Darwin, T. L. Arbogast, and J. G. Kidd, "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," *J. Acoust. Soc. Am.*, vol. 117, pp. 292-304, 2005.

[10] D. S. Brungart and W. R. Rabinowitz, "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 106, pp. 1465-1479, 1999.