The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal

Douglas S. Brungart[a)]
Air Force Research Laboratory, 2610 Seventh Street, Wright-Patterson AFB, Ohio 45433-7901

Brian D. Simpson
Veridian, 5200 Springfield Pike, Suite 200, Dayton, Ohio 45431

(Received 24 October 2000; revised 10 May 2002; accepted 10 May 2002)

Although many studies have shown that intelligibility improves when a speech signal and an interfering sound source are spatially separated in azimuth, little is known about the effect that spatial separation in distance has on the perception of competing sound sources near the head. In this experiment, head-related transfer functions (HRTFs) were used to process stimuli in order to simulate a target talker and a masking sound located at different distances along the listener's interaural axis. One of the signals was always presented at a distance of 1 m, and the other signal was presented 1 m, 25 cm, or 12 cm from the center of the listener's head. The results show that distance separation has very different effects on speech segregation for different types of maskers. When speech-shaped noise was used as the masker, most of the intelligibility advantages of spatial separation could be accounted for by spectral differences in the target and masking signals at the ear with the higher signal-to-noise ratio (SNR). When a same-sex talker was used as the masker, the intelligibility advantages of spatial separation in distance were dominated by binaural effects that produced the same performance improvements as a 4–5-dB increase in the SNR of a diotic stimulus. These results suggest that distance-dependent changes in the interaural difference cues of nearby sources play a much larger role in the reduction of the informational masking produced by an interfering speech signal than in the reduction of the energetic masking produced by an interfering noise source. [DOI: 10.1121/1.1490592]

PACS numbers: 43.66.Pn, 43.66.Qp, 43.66.Rq [LRB]

I. INTRODUCTION

In multitalker speech-perception tasks, performance is much better when the target speech signal and the interfering sound sources are located at different azimuth positions in the horizontal plane than when both the target and masking sounds originate from the same location in space. This so-called ''cocktail-party'' phenomenon has been studied extensively with speech maskers (Drullman and Bronkhorst, 2000; Duquesnoy, 1983; Freyman *et al.*, 1999; Hawley *et al.*, 1999; Festen and Plomp, 1990; Peissig and Kollmeier, 1997; Plomp, 1976) and speechlike noise maskers (Bronkhorst and Plomp, 1988, 1992; Plomp and Mimpen, 1979), and these studies have consistently shown that the intelligibility of the target speech increases systematically with the angular separation between the target and the masker. The release from masking can exceed 10 dB when the target is presented directly in front of the listener and the masker is presented near 90 degrees in azimuth (Bronkhorst, 2000).

Several different mechanisms contribute to this improvement in intelligibility. Perhaps the most important is the increase in signal-to-noise ratio (SNR) that inevitably occurs at one of the two ears when the target and masker signals originate from different directions in the horizontal plane. When two competing sources are located at different angles in the horizontal plane, differences in the head-shadowing effects for the two sources will cause one source to have a higher SNR in the left ear than it does in the right ear and the other source to have a higher SNR in the right ear than it does in the left ear. By selectively attending to the ear with the higher SNR (the ''better'' ear), the listener is able to effectively increase the SNR of either of the two sources. Differences in the spectral shapes of the target and masker signals at the better ear, which are determined by the head-related transfer functions (HRTFs) associated with the target and masker locations, can also influence performance (Zurek, 1993). These differences in the relative levels and spectral shapes of the target and masker signals at the better ear can account for most, but not all, of the intelligibility improvement afforded by spatial separation. Spatial unmasking is also influenced by a binaural interaction effect that is based on differences between the low-frequency interaural time delays (ITDs) and interaural level differences (ILDs) of the target and masker signals (Zurek, 1993; Levitt and Rabiner, 1967). Bronkhorst and Plomp (1988) found that the ITD portion of this binaural interaction effect could account for as much as a 5-dB release from masking for a speech source at 0 degrees and a noise masker near 90 degrees when the head-shadow was removed from the stimulus, but that it contributed only about 2.5 dB to the overall spatial release from masking in natural listening where the head-shadow cues were also available. More recently, Zurek (1993) integrated the better-ear and binaural interaction effects into a single model capable of predicting intelligibility with a spatially separated speech target and noise masker. For a more detailed review of the effects of angular separation in the

[a)]Electronic mail: douglas.brungart@wpafb.af.mil ‘‘cocktail-party’’ effect, see the recent reviews by Ericson and McKinley (1997) and Bronkhorst (2000).

One aspect of the ‘‘cocktail-party’’ phenomenon that has received almost no attention in the literature is the role that spatial separation in distance plays in the perception of multiple competing talkers in the region near the listener’s head. Virtually all previous multitalker experiments have focused on relatively distant sound sources, located 1 m or more from the listener. Because the anechoic HRTF is independent of distance in this region (Brungart and Rabinowitz, 1999), differences in the distances of a target and a masker should have no impact on speech intelligibility when their overall levels are similar at the location of the listener. However, when the source is located within 1 m of the head, the HRTF is highly dependent on distance. Specifically, the ILD increases dramatically with decreasing distance in this region, while the ITD increases only modestly (Brungart and Rabinowitz, 1999). There are also substantial distance-dependent spectral changes in the HRTFs of nearby sound sources. Experiments have shown that listeners are able to use these distance-dependent changes in the HRTF to make reasonably accurate judgments about the distances of nearby sound sources in free-field environments (Brungart *et al.*, 1999; Brungart, 1999a). Until recently, however, almost nothing was known about the impact of these distance-dependent changes in the HRTF on the segregation of sound sources near the listener’s head.

In order to examine the effects of distance on the segregation of nearby sources, Shinn-Cunningham and her colleagues (Shinn-Cunningham *et al.*, 2001) have recently adapted Zurek’s model (1993) to account for the effects of spatial separation in distance on the intelligibility of a nearby speech signal masked by a nearby speech-shaped noise source. Their results have shown that virtually all of the effects of spatial separation with a noise masker can be explained by spectral differences in the target and masker signals at the ear with the better SNR, and that binaural factors can explain only 1–2 dB of the release from masking obtained by spatially separating the signal and masker in distance. However, there is some reason to believe that these results may underestimate the advantages of spatially separating multiple speech signals near the head. Recent studies with sound sources at distances greater than 1 m have shown that the binaural cues play a much larger role in the segregation of speech from a competing speech signal at a different location in azimuth than in the segregation of speech from a competing noise signal at a different azimuth (Freyman *et al.*, 1999; Hawley *et al.*, 2000). This difference seems to occur because interfering speech signals and interfering noise signals produce different kinds of masking: interfering noise signals produce only ‘‘energetic’’ masking, while interfering speech signals may produce both ‘‘energetic’’ and ‘‘informational’’ masking (Brungart, 2001b; Freyman *et al.*, 1999; Kidd *et al.*, 1998). In this context, energetic masking refers to the traditional concept of masking where the interfering signal overlaps in time and frequency with the target signal in such a way that portions of the target signal are rendered inaudible. Informational masking refers to the interference that occurs when the target and masker signals do not overlap in time and frequency but the listener is still unable to segregate the acoustic elements of the target signal from the acoustic elements of a similar-sounding masker. Freyman and his colleagues have suggested that listeners derive a greater benefit when two speech signals are spatially separated in azimuth than when a speech signal and a noise signal are spatially separated in azimuth because the listeners are able to use differences in the apparent locations of the two sounds to reduce the informational component of speech-on-speech masking (Freyman *et al.*, 1999; Freyman *et al.*, 2001). If this hypothesis is true, then there is reason to believe that spatial separations in distance that cause a difference in the apparent locations of the target and masking sounds will also produce a greater benefit when a target speech signal is masked by another speech signal than when it is masked by a noise signal. In this experiment, stimuli from a speech corpus that produces primarily informational masking in two-talker listening were used to determine whether the larger binaural advantages that have been reported in the segregation of competing speech signals that are spatially separated in azimuth also occur in the segregation of competing speech signals that are spatially separated in distance near the listener’s head.

II. METHODS

A. Listeners

A total of nine paid listeners, five male and four female, participated in the experiment. All had normal hearing (<15 dB HL from 500 Hz to 8 kHz), and their ages ranged from 21 to 55 years. All of the listeners had participated in previous experiments that utilized the speech materials used in this study.

B. Stimuli

1. Speech materials

The speech stimuli were taken from the publicly available Coordinate Response Measure (CRM) speech corpus for multitalker communications research (Bolia *et al.*, 2000). This corpus consists of phrases of the form ‘‘Ready (call sign) go to (color) (number) now’’ spoken with all possible combinations of eight call signs (‘‘arrow,’’ ‘‘baron,’’ ‘‘charlie,’’ ‘‘eagle,’’ ‘‘hopper,’’ ‘‘laker,’’ ‘‘ringo,’’ ‘‘tiger’’), four colors (‘‘blue,’’ ‘‘green,’’ ‘‘red,’’ ‘‘white’’), and eight numbers (1–8). Thus, a typical utterance in the corpus would be ‘‘Ready baron go to blue five now.’’ Eight talkers (four male, four female) were used to record each of the 256 possible phrases, so a total of 2048 phrases are available in the corpus. Variations in speaking rate were minimized by instructing the talkers to match the pace of an example CRM phrase that was played prior to each recording. The sentences in the corpus, which are band-limited to 8 kHz, were resampled from the original 40-kHz sampling rate to 25 kHz to reduce computation time in the processing of the stimuli. The phrases were time aligned to ensure that the word ‘‘ready’’ started at the same time in all the speech signals in the stimulus, but no additional efforts were made to synchronize the call signs, colors, and numbers in the competing CRM phrases.

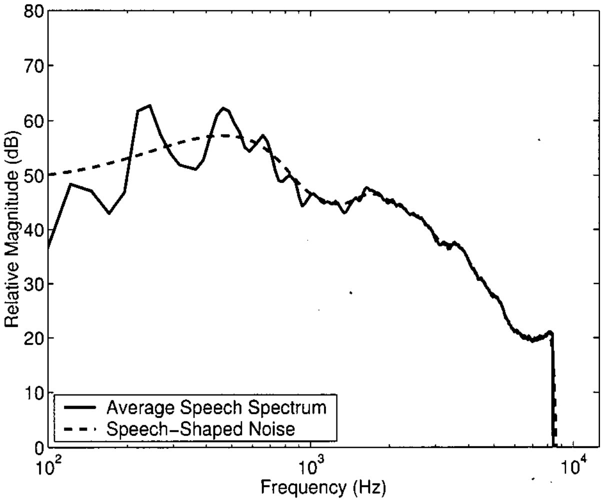


FIG. 1. Average spectrum of the speech utterances in the CRM corpus and frequency response of the filter used to shape the speech-shaped noise maskers. Note that the speech signals in the CRM corpus have been low-pass filtered with an 8-kHz cutoff frequency.

The CRM corpus was selected for this experiment for two reasons. First, the presence of the call sign provides a convenient way to instruct the listener which phrase to attend to in the speech-on-speech masking conditions of the experiment. Second, the small response set of the CRM corpus makes it easy to determine the correct color and number in the target phrase in the presence of relatively high levels of energetic masking (Brungart, 2001a). Previous experiments have shown that this insensitivity to energetic masking causes informational masking to dominate in speech-on-speech masking with the CRM corpus—in two-talker stimuli, listeners are generally able to hear the colors and numbers spoken by both the target and masking talkers, but are unable to correctly determine which color and number were spoken by the target talker (Brungart, 2001b). Thus, the CRM corpus is well suited to experiments such as this one that are designed to concentrate on the informational component of speech-on-speech masking rather than on the energetic component of speech-on-speech masking. Note that one would expect to find smaller differences between speech-on-speech masking and speech-on-noise masking with a measure of speech intelligibility that is more sensitive to the effects of energetic masking, such as the identification of nonsense syllables.

2. Speech-shaped noise

In some trials, a speech-shaped noise signal was used as the masker. The spectrum of this noise masker was determined by averaging the log-magnitude spectra of all of the phrases in the CRM corpus. This average spectrum was used to construct a 129-point finite impulse response (FIR) filter that was used to shape Gaussian noise to match the average spectrum of the speech signals (Fig. 1).

3. Spatial processing

The stimuli in the experiment were processed with HRTFs in order to simulate sound sources at different distances along the listener's interaural axis (Wightman and Kistler, 1989a, b). The HRTFs used for this spatial processing were derived from an earlier set of HRTFs measured for nearby source locations with a Knowles Electronics Manikin for Acoustic Research (KEMAR). These HRTFs, which are described in detail elsewhere (Brungart and Rabinowitz, 1999), were measured in a large anechoic chamber with an acoustic point source located directly to the left of the manikin (90 degrees azimuth) at distances of 12 cm, 25 cm, and 1.0 m from the center of the manikin's head. The overall level effects of distance and the frequency characteristics of the point source were removed from these HRTFs by subtracting the free-field spectrum of the sound source (as measured by a single microphone placed at a location corresponding to the center of the manikin's head) from the HRTFs measured at the manikin's left and right ears. The HRTF measurements were made in the frequency domain and consisted of 600-point transfer functions with 32-Hz resolution from 100 Hz to 19.2 kHz.

The filters used to spatially process the stimuli in this experiment were derived directly from these HRTFs using the following procedure. First, the headphones used in the experiment (Sennheiser HD540) were placed on the KEMAR manikin and the same frequency-domain method used to measure the original HRTFs was used to measure the 600-point left- and right-ear transfer functions of the headphones. These transfer functions were subtracted from the raw HRTFs for the left and right ears in order to determine the desired transfer functions of the headphone-corrected HRTFs for each stimulus location. Then the MATLAB FIR2 command was used to generate 251-point, linear-phase FIR filters matching the magnitudes of the frequency responses of the desired transfer functions over the frequency range from 100 Hz to 15 kHz at a 44.1-kHz sampling rate. These linear-phase filters were up-sampled to a 1-MHz sampling rate in order to delay the contralateral-ear HRTF by the interaural time delay, which was determined from the average slope of the unwrapped phase of the original interaural HRTF over the frequency range from 160 to 1700 Hz.[1] Finally, the HRTFs were down-sampled to a 25-kHz sampling rate to efficiently accommodate the 8-kHz band-limited speech corpus used in this experiment. The resulting HRTFs were stored in a MATLAB file and directly convolved with the target and masker signals immediately prior to each stimulus presentation. Figure 2 shows the frequency responses of the HRTFs used for each source location in this experiment (without headphone correction), and the ILDs and ITDs corresponding to each set of HRTFs.

Although these KEMAR HRTFs do not capture the high-frequency, listener-specific detail that would be present in individualized HRTFs, they do produce distance- and direction-dependent cues that are similar to the ones that would occur with a nearby sound source in the free field. They are therefore able to provide listeners with some information about the directions and distances of virtual sounds. An earlier experiment that required listeners to localize noise bursts that were processed with the same HRTFs used in this experiment has shown that listeners are able to localize both the distances and directions of nearby virtual sounds processed with the KEMAR HRTFs (Brungart and Simpson,

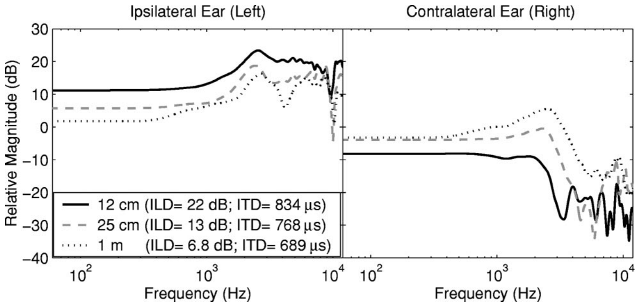


FIG. 2. These curves show the frequency responses of the HRTF filters used to spatially process the stimuli used in the experiment. The headphone response corrections described in the text have been removed from these plots, so they represent the frequency responses of the raw HRTFs measured directly from the KEMAR manikin (as described in Brungart and Rabinowitz, 1999). The numbers in the legend show the average interaural level difference (ILD) (measured from overall rms power for a speech-shaped noise stimulus) and the interaural time delay (ITD) (implemented with a linear phase delay in the HRTF for the contralateral ear) for each stimulus distance used in the experiment. Note that in each case the HRTF has been normalized to the sound pressure level that would occur at the location of the center of the head if the manikin's head were removed.

2001). The polar plot in Fig. 3 shows the median response locations in that experiment for three virtual sound locations along the listener's interaural axis. Although the localization judgments of the listeners in the virtual experiment were generally not as accurate as those of a different group of listeners who were asked to localize a nearby acoustic point source in the free field[2] (Brungart, 1999b), the median response locations shown in the figure indicate that the HRTF processing techniques used in this experiment can be used to generate virtual sounds along the interaural axis that are perceived at systematically increasing distances in roughly the same direction relative to the listener. It is not possible to know exactly what effect the nonindividualized HRTFs used in this experiment had on performance, but it should be noted that previous researchers who have compared the effect of spatial separation in azimuth on multitalker speech perception with virtual sources generated with nonindividualized HRTFs to the effects of spatial separation in azimuth on multitalker speech perception with free-field sources (Nelson *et al.*, 1999; Abouchacra *et al.*, 1997; Hawley *et al.*, 1999) or virtual sources generated with individualized HRTFs (Drullman and Bronkhorst, 2000) have reported no significant differences between the generic virtual presentations and the more realistic free-field and individualized virtual presentations.

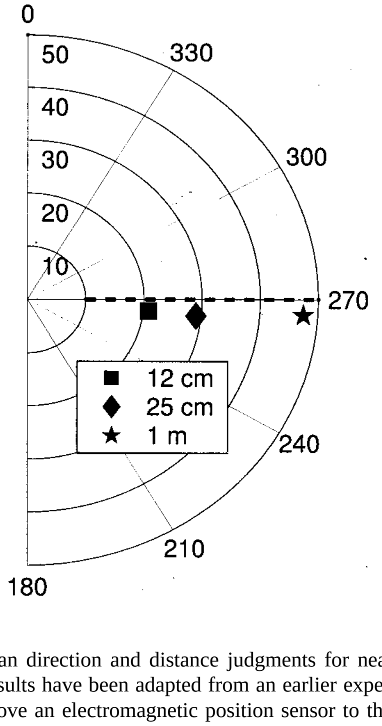


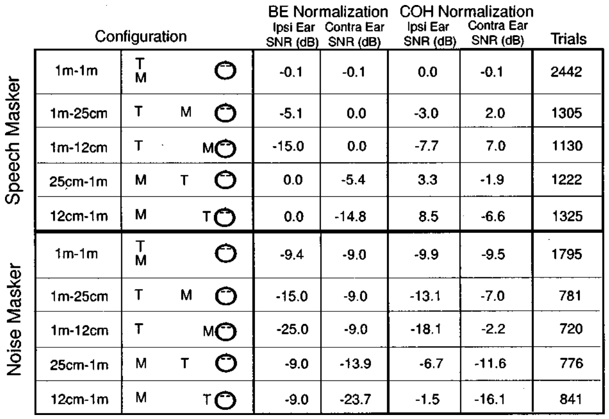
FIG. 3. Median direction and distance judgments for nearby virtual noise bursts. The results have been adapted from an earlier experiment that asked listeners to move an electromagnetic position sensor to the perceived location of a random-amplitude noise burst that was processed with the same set of HRTFs used in this experiment and presented over headphones (Brungart and Simpson, 2001). Each point represents the median location of 162 trials collected with seven normal-hearing listeners: the radius on the polar plot represents the median response distance (in cm), and the angle on the polar plot represents the median response azimuth (in degrees). Although the distance judgments were somewhat compressed, the results clearly show that the stimuli were perceived at systematically increasing distances at approximately the same angle in azimuth.

4. Stimulus configurations

All of the target and masker stimuli were presented along the interaural axis directly to the left of the listener. A total of five different target and masker configurations were tested (as shown in the first two columns of Fig. 4). In the 1 m–1 m configuration, both the target and the masker were presented at the same distance. In the 12 cm–1 m and 25 cm–1 m configurations, the target was presented at a closer distance than the masker. In the 1 m –12 cm and the 1 m–25 cm configurations, the masker was presented at a closer distance than the target. The target and masker locations were selected randomly in each trial in a process that resulted in roughly twice as many trials with the target and masker colocated at 1 m than in the other possible configurations.

5. Normalization

In real-world environments, the overall intensity of a stimulus varies with the distance of the source. Thus, if two equally intense speech signals were separated in distance, one would expect the closer speech signal to be substantially easier to comprehend simply because it would be more intense at the location of the listener; the contribution of binaural cues to the release from masking would be minimal relative to these distance-dependent intensity cues. Therefore, in order to examine the contribution of binaural cues and control for these distance-based intensity variations, the relative levels of the target and masker signals were adjusted in two different ways. In the *center-of-the-head* normalization condition (COH), the overall rms levels of the target and masker signals were equalized before they were convolved with the HRTFs, and the SNRs in the left and right ears were determined by the relative levels of the HRTFs shown in Fig. 2. This effectively normalized the rms levels of the target and masking sounds at the center of the listener's head (with the head removed from the sound field). In contrast, in the *better-ear* normalization condition (BE), the rms levels of target and masker were normalized after they were convolved with the appropriate HRTFs. The SNRs of the spatially processed target and masking signals were computed



	Configuration		BE Normalization		COH Normalization		
			Ipsi Ear SNR (dB)	Contra Ear SNR (dB)	Ipsi Ear SNR (dB)	Contra Ear SNR (dB)	Trials
Speech Masker	1m-1m	T M ○	-0.1	-0.1	0.0	-0.1	2442
	1m-25cm	T M ○	-5.1	0.0	-3.0	2.0	1305
	1m-12cm	T M○	-15.0	0.0	-7.7	7.0	1130
	25cm-1m	M T ○	0.0	-5.4	3.3	-1.9	1222
	12cm-1m	M T○	0.0	-14.8	8.5	-6.6	1325
Noise Masker	1m-1m	T M ○	-9.4	-9.0	-9.9	-9.5	1795
	1m-25cm	T M ○	-15.0	-9.0	-13.1	-7.0	781
	1m-12cm	T M○	-25.0	-9.0	-18.1	-2.2	720
	25cm-1m	M T ○	-9.0	-13.9	-6.7	-11.6	776
	12cm-1m	M T○	-9.0	-23.7	-1.5	-16.1	841

FIG. 4. Target and masker configurations used in the experiment. Column 2 shows a graphical representation of the target and masker locations in each configuration indicated in column 1. Columns 3–6 show the average SNRs (measured from rms power) of the spatially processed target and masker signals in the listener's left (ipsi) and right (contra) ears for better-ear (BE) and center-of-head (COH) normalization. Column 7 shows the number of trials completed in each condition.

from their rms levels at each ear, and the filtered target speech signal was scaled (by an equal amount in both ears) to make the SNR at the ear with the greater SNR (the *better ear*) equal to 0 dB.

The middle columns of Fig. 4 show the average SNRs at the ipsilateral and contralateral ears with BE and COH normalization for the HRTF-processed stimuli in each of the target-masker configurations tested in the experiment. In BE normalization, the SNR at the ear with the higher SNR is forced to be 0 dB, and the SNR in the other ear is determined by the ILDs in the HRTFs. Note that the location of the better ear depends on the relative distances of the target and masker. When the target is closer, the ipsilateral ear is the better ear. When the masker is closer, the contralateral ear is the better ear. In COH normalization, the SNRs at the two ears are determined directly by the normalized levels of the HRTFs shown in Fig. 2. In both BE and COH normalization, the absolute difference between the SNR at the ipsilateral ear and the SNR at the contralateral ear is approximately the same for each target and masker configuration. This difference is approximately equal to the difference in ILD between the HRTF of the target position and the HRTF of the masker position. For example, in the 12 cm–1 m configurations where the ILD is approximately 22 dB for the 12-cm source and 7 dB for the 1-m source (see Fig. 2), the difference between the ipsilateral ear SNR and contralateral ear SNR is approximately 15 dB. Small variations in these average SNR levels occurred because of differences in the spectral content of the target and masking signals and differences in the spectral shapes of the HRTFs at the two ears.

In the speech-shaped noise masker conditions, the masker level was increased by 9 dB after the normalization process in order to produce an SNR of −9 dB at the normalization point. This was done because previous speech-perception experiments in our laboratory have shown that performance with the CRM is most sensitive to changes in the relative level of a speech-shaped noise masker when the SNR of the target phrase is approximately −9 dB (Brungart, 2001b).[3] Note that in the noise-masker conditions shown in Fig. 4, the SNRs in the ipsilateral and contralateral ears are 9 dB lower than in the corresponding configurations with a speech masker.

The signals were presented at a comfortable listening level (approximately 65 dB SPL on average) as measured at the output of the headphones, and the overall level of each stimulus presentation was randomly roved over a 6-dB range (in 1-dB steps). This roving ensured that the listeners were not able to use absolute level to identify the target and masking phrases.

C. Procedure

In each trial, the target phrase was selected randomly from the 256 phrases in the speech corpus with the call sign ‘‘baron,’’ with the restriction that each talker was used the same number of times in each listening session. In the trials with a speech masker, the masking phrase was selected randomly from the 1176 phrases in the speech corpus with a different call sign, a different color coordinate, and a different number coordinate than the target phrase. Note that the random selection of the phrases resulted in same-sex target and masking talkers in 50% of the trials and different-sex target and masking talkers in 50% of the trials. In the trials with a noise masker, a random Gaussian noise was filtered with the speech-shaped noise filter and gated rectangularly to the beginning and end of each phrase. The normalization scheme (COH or BE) was also randomly chosen on each trial.

The data were collected with the listeners seated in front of the CRT of a Windows-based control computer in a quiet, sound-treated listening room. The stimuli for each trial were generated by an interactive MATLAB script, which selected the stimulus signals, processed the signals with the appropriate HRTFs, and presented the signals over headphones (Sennheiser HD540) through a Soundblaster AWE-64 sound card. The listeners were instructed to listen for the target phrase, which was always addressed to the call sign ‘‘baron,’’ and use the mouse to select the color and number contained in the target phrase from an array of colored digits displayed on the screen of the control computer. Each listener first participated in a total of 1560 trials with a speech masker. These trials were collected in 13 blocks of 120 trials each, with each block taking approximately 15 min to complete. Each listener then heard a total of 1000 trials with a speech-shaped noise masker. These trials were collected in five blocks of 200 trials each, with each block taking approximately 20 min to complete. One or two blocks were run per day for each listener over a period of several weeks. Note that some of the data were collected with normalization schemes or target-masker distance configurations that are not discussed in this paper, and that these points were excluded from the data analysis. Thus, the results that follow represent a total of 7435 trials collected with the speech masker and 4913 trials collected with the noise masker.

The distribution of these trials across the different target-masker configurations is shown in the last column of Fig. 4. Within each configuration, approximately half of the trials were conducted with BE normalization and approximately

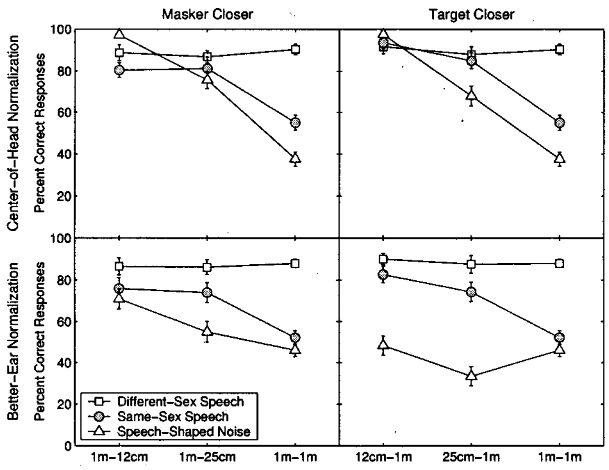


FIG. 5. Percentage of correct color and number identifications for two competing sound sources directly to the left of the listener (90 degrees azimuth). The left panels show performance for each target-masker configuration when the masker is closer than the target. The right panels show performance when the target is closer than the masker. The two rows represent the two different types of normalization used in the experiment. The symbols represent different kinds of target and masking signals (as indicated by the legend). Note that the 1m–1m condition is shown in both columns of the figure. The error bars show 95% confidence intervals calculated from the raw data for each data point.

half of the trials were conducted with COH normalization. Because the number of trials in each possible condition varied across the listeners, all of the mean performance values in the results that follow were calculated by first finding the mean performance values of each listener in that condition and then averaging across these nine individual means to determine overall performance. The standard errors bars shown in each condition represent an rms combination of the nine standard error values calculated for the individual listeners.

III. RESULTS AND DISCUSSION

A. Overall results

The overall results of the experiment are shown in Fig. 5. The data from each target-masker configuration have been plotted separately for BE normalization (bottom row) and COH normalization (top row), and the speech-masking data have been plotted separately for same-sex masking speech (circles), different-sex masking speech (squares), and speech-shaped masking noise (triangles). The left column shows the results for configurations where the masker was closer than the target talker, and the right column shows results for configurations where the target talker was closer than the masker. Each data point in the figure represents the percentage of trials in which the listeners correctly identified both the color and the number contained in the target phrase containing the call sign ''baron,'' and the error bars represent the 95% confidence intervals of each data point. The results indicate that the effects of distance separation on speech intelligibility are different for different types of maskers. When the target speech was masked by a different-sex talker (squares in Fig. 5), spatial separation in distance had little or no impact on performance. The listeners correctly identified both the color and number coordinates in the target phrase in approximately 85%–90% of the trials in all of the conditions tested. Apparently, the monaural cues that allow listeners to segregate different-sex talkers are so effective that no additional intelligibility advantage can be obtained by presenting the target and masking utterances at different distances.

When the target speech was masked by a same-sex talker (circles in Fig. 5), substantial improvements in performance occurred when the target and masking signals were spatially separated in distance. The overall percentage of correct identifications was 30–40 percentage points greater in the 12 cm–1 m configuration than in the 1 m–1 m configuration and about 25 percentage points greater in the 1 m–12 cm configuration than in the 1 m–1 m configuration. There was not much difference between the conditions where the closer talker was at 12 cm and the conditions where the closer talker was at 25 cm: performance was only about 10% better in the 12 cm–1 m configuration than in the 25 cm–1 m configuration (right panels of the figure), and was essentially identical in the 1 m–12 cm and 1 m–25 cm configurations (left panels of the figure). Apparently most of the benefits of spatial separation in distance for same-sex competing talkers can be obtained by moving one of the talkers within 25 cm of the listener's head, even though the 6-dB increase in ILD associated with a decrease in distance from 1 m to 25 cm is much smaller than the 15-dB increase in ILD associated with a decrease in distance from 1 m to 12 cm (Fig. 2).

When the target speech was masked by a speech-shaped noise masker (triangles in Fig. 5), the effects of spatial separation were substantially different for the different normalization conditions. The benefits of spatial separation in distance were greatest with the COH normalization (top panels of the figure), where the percentage of correct identifications systematically increased from approximately 40% in the 1 m–1 m configuration to approximately 75% in the 1 m–25 cm and 25 cm–1 m configurations, and to near 100% in the 1 m–12 cm and 12 cm–1 m configurations. When BE normalization was used (bottom panels of the figure), distance separation had a much smaller effect on performance. When the masker was closer than the target, the percentages of correct responses with BE normalization were 20–25 percentage points lower than in the corresponding configurations with COH normalization, and when the target was closer than the masker, separation in distance essentially had no effect on the intelligibility of the target phrase with BE normalization.

B. Target proximity

The results in Fig. 5 suggest that there are some important differences in performance between conditions where the target was closer than the masker and conditions where the target was farther away. These differences were examined by conducting a three-factor repeated-measures ANOVA for the factors of target-masker configuration (12 cm–1 m or 25 cm–1 m), relative target proximity (target closer or masker closer), and masker type (same-sex speech or speech-shaped noise) with the percentages of correct responses for each listener. The arcsine transform was applied to normalize the percentage data prior to conducting the ANOVA. The results of this ANOVA confirm that there was a significant

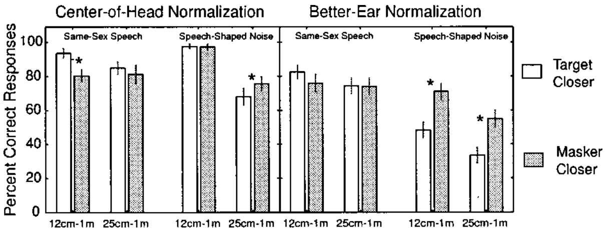


FIG. 6. Comparison of performance when the target was located closer than the masker and when the masker was located closer than the target. The asterisks indicate differences that were significant at the $p < 0.05$ level (two-tailed *t*-tests). The error bars show the 95% confidence intervals calculated from the raw data in each configuration.

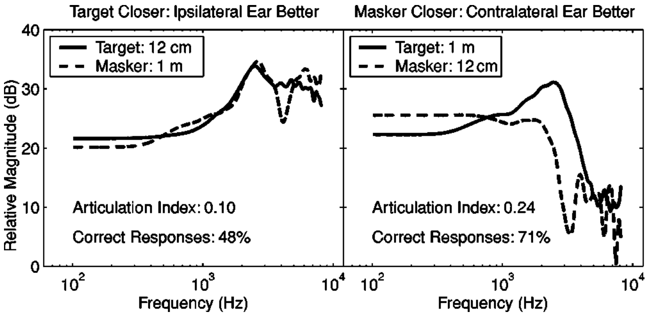


FIG. 7. Effects of spectral shape on the intelligibility of speech at the ear with the better SNR in the 12cm–1m (left panel) and the 1m–12cm (right panel) configurations of the experiment. These are the same 12 cm and 1 m HRTFs shown in Fig. 2, but their relative levels have been adjusted to equalize the overall power of speech-shaped target and masker signals (with the same frequency spectrum shown in Fig. 1) at the location of the better ear with the same method used in the BE normalization conditions of the experiment. In each case, the 20-band method developed by Kryter (1962) has been used to calculate the articulation index (AI) of the target signal when the target is presented at 56 dB SPL and the masker is presented at 65 dB SPL. This difference in AI explains the substantially larger number of correct responses that occurred in the masker-closer conditions of the experiment.

interaction between the relative distance of the target and the masker type ($F_{(1,8)} = 6.573$, $p = 0.033$). This interaction is shown in more detail by Fig. 6, which directly compares performance in the target-closer and masker-closer conditions for each target-masker configuration and each masker type. In the same-sex speech-masking conditions, the relative locations of the target and masker signals had only a modest impact on performance. The only significant difference occurred in the COH condition (left panel of Fig. 6), where performance in the 12 cm–1 m configuration (target closer) was approximately 10% better than in the 1 m–12 cm condition (masker closer). In part, this effect can be explained by the higher SNR that occurred at the better ear in the 12 cm–1 m condition with COH normalization. [Figure 4 shows that the SNR ratio in the better ear was 1.5 dB higher in the 12 cm–1 m configuration than in the 1 m–12 cm configuration (8.5 dB vs 7 dB).] This effect may also reflect a bias on the part of the listeners to direct their attention to the closer talker, who was located only a few centimeters from the ear.

In the noise-masking conditions, the relative distances of the target and masker had the opposite effect on performance. In three of the four noise-masking configurations shown in Fig. 6, performance was significantly worse when the target was closer than the masker than when it was farther away than the masker. This performance differential was particularly large with BE normalization (right panel of the figure), where the percentage of correct responses was more than 20 percentage points higher in the 1 m–12 cm and 1 m–25 cm configurations than in the 12 cm–1 m and 25 cm–1 m configurations. The only noise-masking configuration where performance was not significantly better with a more distant target was the 1 m–12 cm COH configuration, where performance was already near 100% in the 12 cm–1 m configuration and no measurable effect of source proximity was found.

The somewhat counterintuitive effect that relative distance had on performance with the noise masker can be explained by spectral differences in target and masker HRTFs at the ear with the more advantageous SNR in each listening configuration. Figure 7 shows the transfer functions of the HRTFs of the target and masker at the "better ear" for a target at 12 cm and the masker at 1 m (left panel) and for a masker at 12 cm and a target at 1 m (right panel). In both cases, the relative levels of the transfer functions have been normalized with BE normalization to make the overall rms power of a speech-spectrum-shaped noise the same when processed by either HRTF. Note that the spectral shapes of the 12 cm and 1 m HRTFs are substantially more similar at the ipsilateral ear than at the contralateral ear, and that BE normalization produces an SNR in the 2–4-kHz range that is 5–20 dB higher in the masker-closer condition than in the target-closer condition.

In order to analyze the effects of these spectral differences quantitatively, the articulation index (AI) was calculated at the better ear for a speech target presented at 56 dB SPL masked by a speech-shaped noise masker presented at 65 dB SPL with the 20-band method described by Kryter (1962).[4] This calculation indicates that the AI at the better ear was 0.24 when the target was at 1 m and 0.10 when the target was at 12 cm. This difference may explain why performance in the 1 m–12 cm configuration was substantially better than performance in the 12 cm–1 m configuration in the noise-masking conditions with BE normalization. A previous diotic experiment that measured performance in the CRM task as a function of AI with a speech-shaped noise masker (Brungart, 2001a) found that the percentage of correct color-number identifications was approximately 90% when the AI was 0.24 (compared to 71% in the 1 m–12 cm configuration of this experiment) and approximately 50% when the AI was 0.1 (compared to 48% in this 12 cm–1 m configuration of this experiment).

It is interesting to note that the spectral advantages of a more distant target talker did not have any meaningful effect on performance with the speech masker. This can be explained by the fact that speech-on-speech masking with the CRM speech task is dependent primarily on informational masking rather than on energetic masking. Previous experiments have shown that listeners are able to hear both the target and masking phrases with the CRM corpus at SNRs near 0 dB, and that most of their incorrect responses occur because the listeners are unable to segregate the content of the target phrase from the content of the masking phrase (Brungart, 2001b). Thus it is not surprising that relative performance in this task does not correspond to the predictions of the articulation index, which was designed specifically to characterize the effects of energetic masking in speech.

C. Characterizing the advantages of spatial separation in distance

To this point, the performance advantages that occur when two talkers are spatially separated in distance have been described in terms of a difference in the percentage of correct identifications in the CRM task. Although this method of measuring ‘‘spatial advantage’’ is appropriate for comparing performance across different spatialization conditions with the same masking signal (as in the different-sex, same-sex, and speech-shaped noise curves in Fig. 5), it is generally not appropriate for comparing the relative advantages of spatial separation across different speech stimuli or different masking signals. These comparisons require methods of measuring spatial advantage that do not depend on the particular characteristics of the speech intelligibility test used to make the measurements.

One measure of spatial advantage that can be generalized across different speech perception tests is the change in the speech reception threshold (SRT) that occurs when the target and masking signals are spatially separated. The SRT is defined as the minimum presentation level of the target speech required to produce a predetermined threshold level of performance in the speech perception task. It is usually measured by adaptively adjusting the level of the target speech until the desired threshold level of performance is reached. A decibel measure of spatial advantage can be obtained by subtracting the SRT measured in the spatially separated condition from the SRT measured in the nonspatialized condition. This method of measuring spatial advantage does not depend on the difficulty of the particular speech intelligibility task used in the experiment, so it provides a better means of comparing the advantages of spatial separation across different stimulus and masker types than the change in the percentage of correct responses. It also allows direct comparison to the large number of experiments in the literature that have measured spatial advantage this way with noise maskers (Bronkhorst and Plomp, 1988; Festen and Plomp, 1990; Hawley *et al.*, 2000; Shinn-Cunningham *et al.*, 2001; Peissig and Kollmeier, 1997) and with speech maskers (Duquesnoy, 1983; Hawley *et al.*, 2000; Peissig and Kollmeier, 1997).

In this experiment, performance in each condition was measured only at the two SNR values determined by the BE conditions and the COH conditions shown in Fig. 4. Because these SNR values resulted in different levels of performance in each condition tested, it is not possible to directly determine a decibel measure of spatial advantage from these results. It is, however, possible to derive a decibel estimate of spatial advantage by comparing the results of this experiment to the results of a previous experiment that used the same panel of listeners and the same CRM stimuli to measure performance as a function of SNR for nonspatialized (diotic) stimuli (Brungart, 2001b). This method of estimating spatial advantage is illustrated in Fig. 8. The figure separates the data into four separate panels, with each row corresponding to a different target-masker configuration and each column corresponding to a different type of masking sound. Within each panel, the symbols show performance in each target-masker configuration as a function of the SNR at the better ear. The open symbols show performance in the BE conditions (where the better-ear SNR was forced to 0 dB), and the filled symbols show performance in the COH conditions (where the better-ear SNR was determined by the HRTFs). The error bars on each symbol represent the 95% confidence intervals for each data point. The lines in each panel of the figure show performance as a function of SNR from the previous nonspatialized (diotic) experiment that used the same speech-in-speech and speech-in-noise stimuli used in this experiment (Brungart, 2001b).

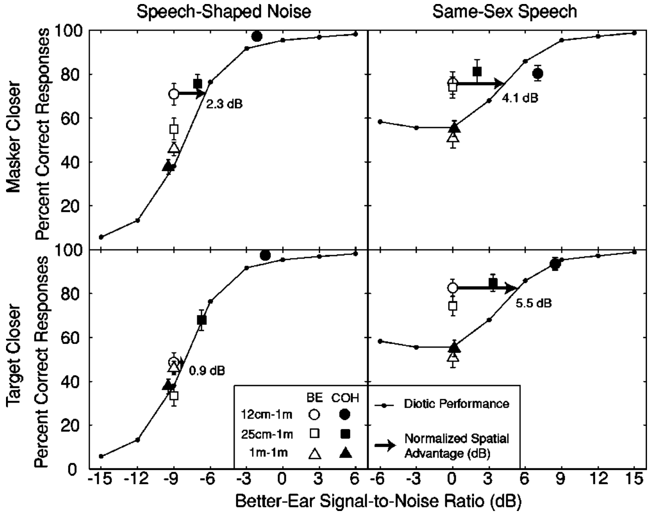


FIG. 8. Percentage of correct color and number identifications for two competing sound sources directly to the left of the listener (90 degrees azimuth) as a function of the SNR at the better ear. The left column shows results for the speech-shaped noise masker, and the right column shows results for the same-sex speech masker. The top row shows results when the masker is closer than the target, and the bottom row shows results when the target is closer than the masker. The results from the 1m–1m configuration are shown in both rows. The symbols show results for the different target-masker configurations in each panel. The open symbols represent conditions with BE normalization, and the shaded symbols represent conditions with COH normalization. The solid line shows performance as a function of signal-to-noise ratio in diotic (nonspatialized) presentations of the same target and masker signals to the same panel of nine listeners (Brungart, 2000b). The arrows in each panel show a decibel estimate of spatial advantage that has been calculated from the difference between the better-ear SNR in the BE 12cm–1m conditions (open circles) and the SNR required to achieve the same level of performance in the corresponding diotic conditions (solid lines in each panel). The error bars show 95% confidence intervals calculated from the raw data for each condition.

In the 1 m–1 m listening configurations, shown by the triangles in Fig. 8, overall performance was approximately the same as in the corresponding diotic configuration with the same SNR value at the better ear. This is not a surprising result, because no binaural or spectral difference cues were available to help segregate the target and masking signals in the 1 m–1 m configurations. It does, however, indicate that the 1 m HRTFs that were used to spatially process the stimuli had little impact on the overall intelligibility of the target speech.

In the spatially-separated listening configurations, performance was generally better than in the corresponding diotic configurations with the same SNR values at the better ear. The right-facing arrows in the figure show a decibel estimate of this spatial advantage for the 12 cm–1 m and 1 m–12 cm listening configurations with BE normalization (open circles in the figure). In each case, the spatial advantage was estimated from the difference between the better-ear SNR in the spatially separated condition and the minimum SNR value required to achieve the same percentage of correct identifications in the corresponding diotic condition. For example, with the same-sex speech masker, correct identifications occurred in approximately 75% of the trials in the 12 cm–1 m configuration when the better-ear SNR was 0 dB (open circle in the top-right panel of the figure). In order to obtain comparable performance in the diotic condition, a better-ear SNR of approximately 4.1 dB would be required (arrow in the figure). Thus, the ''normalized'' spatial advantage in the 12 cm–1 m configuration is about 4.1 dB. Note that we refer to this estimate of spatial advantage in the BE condition as ''normalized'' spatial advantage because it compares performance in nonspatialized and spatialized listening configurations that produce the same SNR at the better ear: this is in contrast to other measures of spatial advantage that include the effects of any increase in better-ear SNR in the spatially separated condition.

These estimates of normalized spatial advantage clearly illustrate the differences that can occur between the percentage estimates of spatial advantage and decibel estimates of spatial advantage. Spatial separation produced the largest percentage point increase in performance (≈30%) in the 12 cm–1 m speech-shaped noise condition with a closer masker (upper left corner of Fig. 8). However, because performance in the CRM test increases much faster with SNR with a noise masker than with a speech masker, the decibel spatial advantage was substantially larger for the speech-masking conditions than for the noise-masking conditions.

The larger normalized spatial advantages that occurred in the speech-masking conditions of the experiment also suggest that binaural difference cues play a greater role in distance-based speech segregation with a speech masker than with a noise masker. The overall decibel measure of normalized spatial advantage shown in Fig. 8 includes the effects of two different types of spatial segregation cues. The first is a monaural spectral cue based on differences in the spectral characteristics of the target and masking signals at the better ear. The second is a binaural cue, often referred to as binaural advantage or binaural interaction (Zurek, 1993; Hawley *et al.*, 2000), which allows listeners to segregate sounds on the basis of variations in the interaural difference cues produced by the target and masking signals. In this experiment, there is reason to believe the monaural spectral cue was responsible for most of the 2.3-dB normalized spatial advantage found in the 1 m–12 cm noise configuration shown in the top left panel of Fig. 8. As discussed in the previous section, differences in the spectral shapes of the 12 cm and 1 m HRTFs produce an SNR advantage at the better ear that substantially improves performance with the noise masker in the masker-closer configurations but has little effect on performance in the target-closer configurations. These monaural spectral cues could explain most of the 2.3-dB spatial advantage found in the masker-closer configurations, and the absence of these spectral cues may explain why the spatial advantage was smaller (0.9 dB) in the target-closer configurations. At the same time, there is little evidence that monaural spectral cues had much influence on performance in the speech-masker conditions. Indeed, the spatial advantage was actually lower in the masker-closer conditions where the monaural spectral segregation cues should have provided the most benefit. Overall, though, the spatial advantage was roughly comparable in the target-closer and masker-closer configurations with the speech masker, suggesting that the segregation was based primarily on binaural difference cues that were symmetric across the two configurations. Thus, it appears that binaural cues were responsible for a large portion of the relatively large spatial advantages found for the same-sex speech masker, but only for a small portion of the relatively small spatial advantages found with the speech-shaped noise masker.

Further evidence that the segregation of speech from a noise masker is dominated by monaural cues is provided by the relatively large influence that better-ear SNR had on performance in the noise-masking conditions of the experiment. In all the spatially separated configurations tested with the noise masker, large increases in performance occurred with the larger better-ear SNR values in the COH normalization conditions (comparing the open and black symbols in the left column of Fig. 8). This is in direct contrast to the spatially separated configurations with the speech masker, where performance increased only modestly in the COH conditions (comparing the open and black symbols in the right column of Fig. 8). Clearly the SNR at the more advantageous ear was a more important factor in determining performance with the noise masker than it was with the speech masker.

Overall, these data suggest that listeners who are attempting to segregate a speech signal from a speech masker perform much better when they have access to the acoustic signals at both ears than when they only have access to the acoustic signal at the ear with the higher SNR. In contrast, listeners who are attempting to extract information from a speech signal masked by noise receive little benefit from having access to the signals at both ears. This supports the hypothesis that the binaural difference cues associated with spatial separation in distance contribute more to spatial unmasking with an informational speech masker than with an energetic noise masker. It does not, however, make it clear whether the difference is due to low-level binaural signal processing or if it is a higher-level process related to a difference in the apparent locations of the sounds. The next section describes a second experiment that was designed to explore this issue in more detail.

IV. DISTANCE PERCEPTION IN TWO-TALKER SPEECH STIMULI

The results of the first experiment suggest that listeners are able to use the distance-dependent changes that occur in the HRTFs of nearby sound sources to segregate speech signals that originate from different locations along the interaural axis. They do not, however, provide any information about where the listeners perceived those speech signals. In order to draw any definitive conclusions about the effect that spatial separation in distance has on multitalker speech segregation, it is necessary to verify that two conditions were met by the virtual stimuli presented in this experiment: (1) that the target and masking signals appeared to be located at the same angle in azimuth and (2) that the target and masking signals appeared to be located at different distances.

The first of these conditions is required to ensure that the listeners were not performing the segregation task on the basis of differences in the apparent directions of the target and masking signals. The results of a previous experiment that measured localization judgments for noise bursts that were processed with the same set of HRTFs used in experiment 1 (Brungart and Simpson, 2001) provide some evidence that the two-talker stimuli used in this experiment met this ‘‘equal apparent azimuth’’ requirement. These results, which are illustrated in Fig. 3, show that 12 cm, 25 cm, and 1 m HRTFs measured at 270 degrees in azimuth with the same techniques used to produce the HRTFs used in this experiment all resulted in median location judgments near 270 degrees. Although there was some variability in the responses, these results show that there was no systematic tendency to perceive the stimuli at different locations in azimuth. Even if there were small differences in the apparent azimuth locations of the signals, the impact of these differences would be limited somewhat by the relative insensitivity of human listeners to changes in the directions of sounds near the interaural axis. Experiments that have measured the minimum audible angle (MAA) for lateral source positions have found that the MAA is roughly 10 degrees for sound sources at 75 degrees azimuth and roughly 20 degrees for sound sources at 90 degrees azimuth (Chandler and Grantham, 1992; McKinley *et al.*, 1994). Thus, the 3–6-degree variations in the median azimuth judgments for the different HRTF distances shown in Fig. 3 are small relative to the MAA in this region. Consequently, we do not believe the advantages of spatial separation in distance found in the first experiment can be explained by differences in the apparent azimuth locations of the competing signals.

The results shown in Fig. 3 also provide evidence that the ‘‘different apparent distance’’ requirement was met by the stimuli used in the first experiment. Although the responses were compressed relative to the range of simulated distances, the median response distances increased systematically with the simulated distances of the noise bursts. However, this experiment did not measure perceived distance with virtual speech sounds, and it did not measure the listener's ability to localize the distances of two simultaneous virtual sounds. This makes it difficult to know for certain whether the listeners in the first experiment were actually perceiving the target and masking signals at different distances. In order to address this issue, a second experiment was conducted that examined how well listeners were able to judge the relative distances of the target and masking talkers in the two-talker stimuli used in the first experiment.

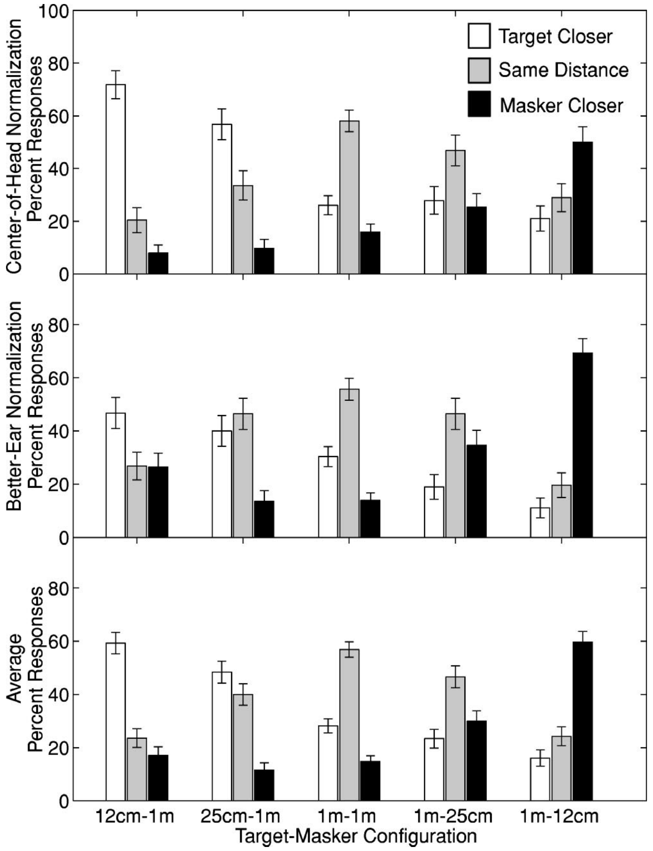


FIG. 9. Distribution of responses in experiment 2. Each group of three bars shows the percentages of target-closer, same-distance, and masker-closer responses for a different target-masker configuration. Note that the leftmost bar in each group represents correct responses in the 12 cm–1 m and 25 cm–1 m configurations, the middle bar represents correct responses in the 1 m–1 m configurations, and the rightmost bar represents correct responses in the 1 m–25 cm and 1 m–12 cm configurations. The top panel shows the results for COH normalization, the middle panel shows the results for BE normalization, and the bottom panel shows the results averaged across both normalization conditions. The error bars show 95% confidence intervals calculated from the raw data in each configuration.

A. Methods

The stimuli used in experiment 2 were almost identical to those used in the speech-masking conditions of experiment 1. They consisted of pairs of randomly selected phrases from the CRM corpus that were processed with KEMAR HRTFs measured at different distances (12 cm, 25 cm, or 1 m) along the listener's interaural axis. One of the phrases (the target phrase) always contained the call sign ‘‘baron,’’ and the other phrase (the masking phrase) always contained the call sign ‘‘ringo.’’ The HRTF-processed stimuli were normalized to have an SNR equal to 0 dB either at the center of the head (COH normalization) or at the ear with a higher SNR (BE normalization). Then they were mixed together digitally and played back to the listener over stereo headphones (Sennheiser HD540) in a quiet, sound-treated listening room. The overall level of each stimulus presentation was randomly roved over a 6-dB range in 1-dB steps.

Although the stimuli were similar to those used in the first experiment, the task was quite different. After each stimulus presentation, the listeners were asked to determine whether the target phrase was closer than the masking phrase, the same distance as the masking phrase, or farther away than the masking phrase. They responded by using the computer mouse to select ‘‘Baron Closer,’’ ‘‘Same Distance,’’ or ‘‘Baron Farther’’ on the screen of the control computer. No feedback was provided about the actual locations of the stimuli.

A total of seven listeners participated in the second experiment, including five who also participated in the first experiment. Each listener participated in four blocks of 120 trials, with each block consisting of five repetitions of all combinations of two target-masker voice configurations (same-sex or different-sex), two normalization schemes (BE or COH), and five target-masker distance configurations (12 cm–1 m; 25 cm–1 m; 1 m–1 m; 1 m–25 cm; and 1 m–12 cm). Note that, as in the first experiment, twice as many trials were collected in the 1 m–1 m configuration than in the other configurations.

B. Results and discussion

Figure 9 shows the distribution of responses for each target-masker configuration and normalization scheme used in the experiment. The top panel shows the results for COH normalization, the middle panel shows results for BE normalization, and the bottom panel shows the results averaged across these two normalization schemes. Within each target-masker configuration, the left bar shows the percentage of target-closer responses, the middle bar shows the percentage of same-distance responses, and the right bar shows the percentage of target-farther responses (as indicated in the legend). The error bars represent the 95% confidence intervals of each data point.

The averaged data shown in the bottom panel of the figure indicate that the listeners were able to make reasonably accurate judgments about the relative distances of the target and masking phrases. In the 12 cm–1 m, 1 m–12 cm, and 1 m –1 m configurations, the listeners correctly identified the relative location of the target phrase in approximately 60% of the trials. While this is far from perfect performance, it is well above chance and is perhaps remarkably good when one considers that the task required the listener to correctly identify the phrase containing ‘‘baron’’ while simultaneously determining the location of that phrase relative to the masker. It is also important to note that these results cannot be explained by distance-dependent differences in the overall levels of the competing talkers. Although there was a systematic relation between overall intensity and distance in the COH normalization conditions (where the closer talker was always more intense in the left ear and less intense in the right ear), there were no consistent level-based distance cues in the BE normalization conditions. In the 12 cm–1 m condition with BE normalization, for example, the closer target talker was presented at the same level as the more distant masking talker in the left ear, and at a level almost 15 dB lower than the more distant masking talker in the right ear (see Fig. 4). Thus, depending on how the listeners integrated the intensity of the stimuli across the two ears, the overall level cue in the 12 cm–1 m condition with BE normalization was either nonexistent or in opposition to the usual inverse relationship between intensity and distance. Despite this misleading intensity cue, the listeners in this condition performed well above chance in identifying the 12 cm target talker as the closer talker (left-hand side of the middle panel in Fig. 9). The ability of the listeners to correctly identify the relative distances of the target and masking talkers without feedback and in the presence of misleading intensity cues confirms that the KEMAR HRTFs used in these experiments provided sufficiently realistic acoustic cues for the listeners to perceive the competing talkers at different distances.

It is, however, apparent that the listeners were much better at segregating the two speech messages in experiment 1 than they were at determining the relative distances of the two speech signals in experiment 2. In the 12 cm–1 m configuration with BE normalization, for example, the listeners correctly identified both the color and number in the target phrase approximately 85% of the time in experiment 1 (lower right panel of Fig. 4), but correctly identified the relative distance of the target talker in less than 50% of the trials of experiment 2. Most of this reduction in performance can likely be attributed to the increased complexity of the task in experiment 2, where the listeners had to both identify the target phrase and determine its relative distance at the same time. In contrast, listeners were required only to identify the target phrase in experiment 1.

One final interesting aspect of the data from the second experiment is that there were no indications of the large differences in performance that occurred in the same-sex and different-sex masking conditions of experiment 1. In fact, overall performance was identical with the same- and different-sex masking voices in experiment 2 (52% correct responses). Thus, although both experiments required the listeners to segregate the target phrase from the masking phrase, it is apparent that differences in the vocal characteristics of the two talkers provided a much larger benefit in the speech intelligibility task in experiment 1 than in the distance localization task in experiment 2.

V. DISCUSSION AND CONCLUSIONS

The results of these experiments provide insights into the role that spatial separation in distance plays in determining the intelligibility of a nearby talker masked by a competing nearby sound. When the listening task is relatively easy to perform with spatially co-located signals, spatial separation of the target and masker in distance does not improve the intelligibility of the talker. This is apparent from the lack of any discernible differences between the conditions where the target and masker phrases were presented at different distances and those where they were presented at the same distance when two phrases were spoken by different-sex talkers.

When the target phrase is masked by noise, spatially separating the target and masker can produce a tremendous improvement in speech intelligibility (from 40% to near 100%). However, nearly all of this benefit is derived from spectral differences in the target and masker signals at the better ear. The binaural difference cues that appear to dominate the perception of distance for nearby sound sources (Brungart, 1999a) contribute little or nothing to our ability to segregate a nearby talker from a masking noise—the overall spatial advantage was not much larger than 2 dB even for the largest spatial separations used in this experiment, and most of this overall spatial advantage was the result of monaural spectral cues at the listener's better ear. This is consistent with the findings of Shinn-Cunningham *et al.* (2001), who showed that spatial unmasking effects with a noise masker are dominated by spectral ''better ear'' advantages, and that binaural interaction effects account for less than 2 dB of a total threshold shift of 25 dB or more for sound sources located at 15 cm and 1 m along the interaural axis of the listener. They also found that their predictions overestimated performance when the target and masker were at 90 degrees, suggesting that even 2 dB is a generous estimate of the actual binaural advantage that can be obtained by separating the distances of a nearby speech signal and noise masker.

When the target phrase is masked by same-sex speech, however, spatially separating the target and masker can produce improvements in intelligibility that substantially exceed those that would be predicted from spectral differences at the better ear. The overall spatial advantage of separating the target and masker in distance was as large as 5.5 dB when the SNR at the better ear was normalized to 0 dB, and monaural spectral cues seemed to contribute very little to this overall spatial advantage. In fact, in contrast to the relatively minor role that binaural cues play in spatial unmasking with a noise masker, binaural cues appear to account for most of the spatial unmasking that occurs with a speech masker. The results in Fig. 5 show that spatial separation in distance improved performance with a same-sex speech masker by about 28 percentage points with BE normalization and by about 33 percentage points with COH normalization. Thus, about 85% of the intelligibility improvement afforded by spatially separating the talkers in distance was maintained when the SNR advantage at the better ear was eliminated. With the noise masker, spatial separation in distance improved performance by about 60 percentage points with COH normalization and only by about 15 percentage points with BE normalization (averaging the target-closer and masker-closer conditions). Thus, only about 25% of the intelligibility improvement was maintained when the better-ear SNR advantage was eliminated with a noise masker. Clearly binaural difference cues play a much larger role in the spatial unmasking of sound sources that are spatially separated in distance when the masker is same-sex speech than when the masker is speech-shaped noise.

These results are consistent with other recent multitalker experiments that have found similar differences between the binaural advantages of spatial separation in azimuth with a distant noise masker or a distant speech masker. Hawley *et al.* (2000), for example, also found that the binaural advantages of spatial separation in azimuth were 3–4 dB larger when two or more speech or time-reversed speech signals were used as the maskers than when two or more noise or modulated noise signals were used as the maskers. Freyman *et al.* (1999) used the precedence effect to manipulate the apparent locations of a target and a masker without affecting the SNR at either ear, and found that apparent location had a substantial effect on speech intelligibility with a speech masker but essentially no effect with a noise masker. Thus it appears to be generally true that the binaural difference cues associated with spatial separation of the target and masking sounds have a much greater impact on speech intelligibility with speech maskers than with noise maskers. Freyman and his colleagues have suggested that this occurs because differences in the apparent locations of sounds can dramatically reduce informational masking by enhancing the listener's ability to selectively attend to the target speech and avoid being distracted by the contents of the interfering speech signal. If this is the case, then the results of this experiment can be explained in the same way: the informational masking that occurred with the same-sex speech masker was reduced when the target and masker were spatially separated in distance because they appeared to originate at different locations in space. The results of the second experiment support this hypothesis, because they show that the listeners generally perceived the talkers at different distances along the interaural axis. However, further research is necessary to conclusively determine whether the intelligibility advantages of spatial separation found in experiment 1 were caused by differences in the apparent locations of the competing sounds or if they were the result of some other kind of binaural processing related to the enlarged ILDs that occur for sound sources near the head.

Although the mechanisms involved are not yet fully understood, it is clear from the results of these experiments that the spatial unmasking of speech associated with the traditional ''cocktail-party'' effect can be achieved by spatial separation in distance as well as spatial separation in direction when the target and masking sounds are located near the listener. Additional experiments are now needed to examine the effects that spatial separation in distance has at locations off of the listener's interaural axis in order to form a more complete picture of the interactions that occur between apparent distance and apparent direction in the segregation of competing sound sources near the head.

ACKNOWLEDGMENTS

The authors would like to thank Rich Freyman and Larry Feth for their insightful comments during the preparation of the manuscript. Portions of this research were presented at the 2001 International Conference on Auditory Display in Espoo, Finland. This research was funded in part by AFOSR LRIR 01HE01COR.

[1] Although this does not preserve the exact phase information of the original HRTFs, previous research has shown that linear-phase HRTFs that maintain the correct low-frequency phase information are indistinguishable from HRTFs that preserve the original phase information (Kulkarni *et al.*, 1999).

[2] The absolute azimuth errors were approximately 16 degrees with the virtual sounds and approximately 9 degrees with the free-field sounds. The stimulus-response correlation coefficient in distance for sources along the interaural axis was approximately 0.6 for the virtual sounds and approximately 0.8 for the free-field sounds.

[3] If the experiment were conducted with a 0 dB SNR in the better ear with a noise masker, performance in the CRM task would asymptote to near 100%. Conversely, if the experiment were conducted with a −9 dB SNR in the better ear with a speech masker, performance would be in a region where SNR is known to have relatively little effect on performance. In order to avoid these problems, the noise masker was normalized to a level 9 dB higher than the level of the speech masker in each corresponding stimulus configuration.

[4] Note that this analysis assumed a speech-shaped noise signal that exactly matched the spectrum of the talker. This differs slightly from the actual noise masking conditions of the experiment, where the speech-shaped noise signal matched the average spectrum across all the talkers in the corpus and not the average spectrum of any individual talker. Presumably the AI would be slightly higher with a noise masker that did not exactly match the spectrum of the target talker. This is likely to be a small effect, however. Festen and Plomp (1990) found only a small difference between speech-shaped noise that matched the spectrum of a same-sex talker and noise that matched the spectrum of a different-sex talker.

Abouchacra, K., Tran, T., Besing, J., and Koehnke, J. (**1997**). ''Performance on a selective attention task as a function of stimulus presentation mode,'' in *Proceedings of the Midwinter Meeting of the Association for Research in Otolaryngology*, St. Petersburg Beach, Florida.

Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (**2000**). ''A speech corpus for multitalker communications research,'' J. Acoust. Soc. Am. **107**, 1065–1066.

Bronkhorst, A. (**2000**). ''The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions,'' Acustica **86**, 117–128.

Bronkhorst, A., and Plomp, R. (**1988**). ''The effect of head-induced interaural time and level difference on speech intelligibility in noise,'' J. Acoust. Soc. Am. **83**, 1508–1516.

Bronkhorst, A., and Plomp, R. (**1992**). ''Effects of multiple speechlike maskers on binaural speech recognition in normal and impaired listening,'' J. Acoust. Soc. Am. **92**, 3132–3139.

Brungart, D. (**1999a**). ''Auditory localization of nearby sources. III: Stimulus effects,'' J. Acoust. Soc. Am. **106**, 3589–3602.

Brungart, D. (**1999b**). ''Auditory Parallax Effects in the HRTF for Nearby Sources,'' in *Proceedings of 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 17–20 October 1999, pp. 171–174.

Brungart, D. (**2001a**). ''Evaluation of speech intelligibility with the coordinate response measure,'' J. Acoust. Soc. Am. **109**, 2276–2279.

Brungart, D. (**2001b**). ''Informational and energetic masking effects in the perception of two simultaneous talkers,'' J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D., and Rabinowitz, W. (**1999**). ''Auditory localization of nearby sources. I: Head-related transfer functions,'' J. Acoust. Soc. Am. **106**, 1465–1479.

Brungart, D., and Simpson, B. (**2001**). ''Auditory Localization of Nearby Sources in a Virtual Audio Display,'' in *Proceedings of 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 21–24 October 2001, pp. 107– 110.

Brungart, D., Durlach, N., and Rabinowitz, W. (**1999**). ''Auditory localization of nearby sources. II: Localization of a broadband source,'' J. Acoust. Soc. Am. **106**, 1956–1968.

Chandler, D. W., and Grantham, D. W. (**1992**). ''Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity,'' J. Acoust. Soc. Am. **91**, 1624–1636.

Drullman, R., and Bronkhorst, A. (**2000**). ''Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation,'' J. Acoust. Soc. Am. **107**, 2224–2235.

Duquesnoy, A. (**1983**). ''Effect of a single interfering noise or speech source on the binaural sentence intelligibility of aged persons,'' J. Acoust. Soc. Am. **74**, 739–943.

Ericson, M., and McKinley, R. (**1997**). ''The intelligibility of multiple talkers spatially separated in noise,'' in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Erlbaum, Hillsdale, NJ), pp. 701–724.

Festen, J., and Plomp, R. (**1990**). ''Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing,'' J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R., Balakrishnan, U., and Helfer, K. (**2001**). ''Spatial release from informational masking in speech recognition,'' J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R., Helfer, K., McCall, D., and Clifton, R. (**1999**). ''The role of perceived spatial separation in the unmasking of speech,'' J. Acoust. Soc. Am. **106**, 3578–3587.

Hawley, M., Litovsky, R., and Colburn, H. (**1999**). ''Speech intelligibility and localization in a multi-source environment,'' J. Acoust. Soc. Am. **105**, 3436–3448.

Hawley, M., Litovsky, R., and Culling, J. (**2000**). ''The 'cocktail party' effect with four kinds of maskers: Speech, time-reversed speech, speech-shaped noise, or modulated speech-shaped noise,'' in *Proceedings of the Midwinter Meeting of the Association for Research in Otolaryngology*, p. 31.

Kidd, G. J., Mason, C., Rohtla, T., and Deliwala, P. (**1998**). ''Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns,'' J. Acoust. Soc. Am. **104**, 422–431.

Kryter, K. (**1962**). ''Methods for calculation and use of the articulation index,'' J. Acoust. Soc. Am. **34**, 1689–1697.

Kulkarni, A., Isabelle, S., and Colburn, H. (**1999**). ''Sensitivity of human subjects to head-related transfer function phase spectra,'' J. Acoust. Soc. Am. **105**, 2821–2840.

Levitt, H., and Rabiner, L. (**1967**). ''Predicting binaural gain in intelligibility and release from masking of speech,'' J. Acoust. Soc. Am. **42**, 820–829.

McKinley, R., Ericson, M., and D'Angelo, W. (**1994**). ''Three dimensional audio displays: development, applications, and performance,'' Aviat. Space Environ. Med. **65**, A31–38.

Nelson, W. T., Bolia, R. S., Ericson, M. A., and McKinley, R. L. (**1999**). ''Spatial audio displays for speech communication. A comparison of free-field and virtual sources,'' in *Proceedings of the 43rd Meeting of the Human Factors and Ergonomics Society*, pp. 1202–1205.

Peissig, J., and Kollmeier, B. (**1997**). ''Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners,'' J. Acoust. Soc. Am. **35**, 1660–1670.

Plomp, R. (**1976**). ''Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of the azimuth of a single competing sound source (speech or noise),'' Acustica **34**, 325–328.

Plomp, R., and Mimpen, A. (**1979**). ''Improving the reliability of testing the speech reception threshold,'' Audiology **18**, 43–52.

Shinn-Cunningham, B., Schickler, J., Kopco, N., and Litovsky, R. (**2001**). ''Spatial unmasking of nearby speech sources in a simulated anechoic environment,'' J. Acoust. Soc. Am. **110**, 1118–1129.

Wightman, F., and Kistler, D. (**1989a**). ''Headphone simulation of free-field listening. I: Stimulus synthesis,'' J. Acoust. Soc. Am. **85**, 858–867.

Wightman, F., and Kistler, D. (**1989b**). ''Headphone simulation of free-field listening. II: Psychological validation,'' J. Acoust. Soc. Am. **85**, 868–878.

Zurek, P. M. (**1993**). ''Binaural advantages and directional effects in speech intelligibility,'' in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. Studebaker and I. Hochberg (Allyn and Bacon, Portland).