THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *14*(3), 313–334

Factors That Influence Intelligibility in Multitalker Speech Displays

Mark A. Ericson, Douglas S. Brungart, and Brian D. Simpson
Air Force Research Laboratory
Human Effectiveness Directorate
Wright-Patterson AFB, Ohio

Although many researchers have commented on the potential of audio display technology to improve intelligibility in multitalker speech communication tasks, no consensus exists on how to design an "optimal" multitalker speech display. In this article, we review several experiments that have used a consistent procedure to evaluate the effect of four monaural parameters on overall intelligibility. We also present the results of a new experiment that has used the same procedure to examine the influence of 2 additional factors in binaural speech displays: (a) the apparent spatial locations of the talkers and (b) the listener's a priori information about the listening task.

Many critically important aviation-related tasks require listeners to monitor and respond to speech messages originating from two or more competing talkers. A classic example of this kind of task occurs in air traffic control in which a controller is required to communicate critical information to and from multiple simultaneous aircraft while maintaining an acute awareness of the relative positions of all the aircraft in their assigned area. Commercial pilots also encounter situations in which they need to communicate with other crew members on the plane and controllers on the ground at the same time. Military pilots face an even more difficult situation in which they may need to communicate with other aircraft in their own formation, command and control personnel in Airborne Warning & Control System aircraft and at ground-based command centers, and ground-based target spot personnel near the site of an air strike. In all of these situations, a well-designed multitalker speech display could improve the overall performance of the operator not only because it may reduce the chances of a potentially deadly miscommunication but

Requests for reprints should be sent to Mark A. Ericson, AFRL/HECB, 2610 Seventh Street, Wright-Patterson AFB, OH 45433–7901. E-mail: mark.ericson@wpafb.af.mil

also because it may reduce the overall workload associated with multitalker listening and allow the operator to attend to other critical tasks.

Many researchers have commented on the substantial benefits that audio display technology can provide in a multitalker communication environment. Some of the earliest efforts in this area used spectral manipulations to enhance the segregation of multiple talkers in a monaural audio channel. For example, in a three-talker system, speech segregation may be enhanced by high-pass filtering one talker, low-pass filtering a second talker, and all-pass filtering the third talker (Spieth, Curtis, & Webster, 1954; U.S. Department of Defense, 1998). More recent efforts have used virtual audio displays to spatially separate the competing speech channels (Begault, 1999; Crispien & Ehrenberg, 1995; Ericson & McKinley, 1997). To this point, however, no consensus has been reached on the design parameters that are most important in determining the effectiveness of multitalker speech displays. In part, at least, this lack of consensus is a result of the extreme complexity of the multitalker listening problem—performance in such tasks depends on a wide variety of factors including (a) the signal-to-noise ratio (SNR) in the communications system, (b) the number of competing talkers, (c) the voice characteristics of the talkers, (d) the relative levels of the talkers, (e) the apparent spatial locations of the talkers, and (f) the listener's a priori knowledge about the listening task. A further complicating issue is the variety of methodologies that have been used to examine these factors; procedural variations often make it difficult to compare the results of different multitalker listening experiments. In this article, we present the results of a number of experiments that have used the Coordinate Response Measure (CRM; Moore, 1981) to examine the impact that different audio display design parameters have on performance in a multitalker communications task. This allows a comparison of the relative importance of each of these parameters that can be used as a guide in the design of multitalker speech displays.

This article is divided into two sections. In the first section, we review a series of experiments that have examined performance in monaural speech displays in which all of the competing talkers were mixed together into a single audio channel prior to presentation to the listener. In the second section, we present new results of an experiment that examined the effects of spatial separation and a priori information in a binaural speech display in which stereo headphones were used to present different audio signals to the listener's two ears.

EXPERIMENTAL METHODOLOGY: THE CRM

All of the experiments described in this article were conducted using the CRM. This speech intelligibility test was originally developed to provide greater operational validity for military communications tasks than standard speech intelligibility tests based on phonetically balanced words. In the CRM task, a listener hears one or more simultaneous phrases of the form “Ready, (Call Sign), go to (color) (number) now” with one of eight call signs (“Baron,” “Charlie,” “Ringo,” “Eagle,” “Arrow,” “Hopper,” “Tiger,” and “Laker”), one of four colors (red, blue, green, and white), and one of eight numbers (1–8). The listener’s task is to listen for the target sentence containing their preassigned call sign (usually “Baron”) and respond by identifying the color and number coordinates contained in that target phrase.

Although the CRM was originally intended to measure speech intelligibility with a noise masker, its call-sign-based structure makes it ideal for use in multitalker listening tasks. The embedded call sign is the only feature that distinguishes the target phrase from the masking phrases, so the listener is forced to attend to the embedded call signs in of all the simultaneous phrases to successfully extract the information contained in the target phrase (Abouchacra, Tran, Besing, & Koehnke, 1997; Spieth et al., 1954). In this regard, it is similar to many command and control tasks in which operators are required to monitor multiple simultaneous channels for important information that may originate from any channel in the system. However, because the simple sentence structure and test words provide no syntactic information to the listener, the CRM may not be representative of performance in all communications tasks.

The experiments we describe in this article were conducted using the corpus of CRM speech materials that has been made publicly available in CD-ROM format by researchers at the Air Force Research Laboratory (Bolia, Nelson, Ericson, & Simpson, 2000). This CRM corpus contains all 256 possible CRM phrases (8 call signs × 4 colors × 8 numbers) spoken by eight different talkers (four male, four female). The experiments we describe in the following sections were conducted using this corpus. In all cases, the stimulus consisted of a combination of a target phrase, which was randomly selected from all of the phrases in the corpus with the call sign “Baron,” and one or more masking phrases, which were randomly selected from the phrases in the corpus with different call signs, colors, and numbers than the target phrase. These stimuli were presented over headphones at a comfortable listening level (approximately 70 dB SPL), and the listener’s responses were collected either by using the computer mouse to select the appropriately colored number from a matrix of colored numbers on the CRT or by pressing an appropriately marked key on a standard computer keyboard. In each of the following sections, we discuss a different factor that influences speech intelligibility in a multitalker listening environment.

MONAURAL FACTORS THAT INFLUENCE MULTITALKER LISTENING

The factors that influence multitalker speech perception can be divided into two broad categories: monaural factors that influence performance in all speech displays and binaural factors that only influence performance in binaural speech displays. In this section, we review the results of experiments that have used the CRM corpus to examine the impact of four monaural factors in multitalker speech perception.

SNR

One factor that influences the performance of any audio display is the overall noise level in the output signal. In the case of a speech display based on radio communications, three different kinds of noise contribute to this overall noise level: (a) ambient noise in the environment of the talker that is picked up by the microphone that records the talker's voice, (b) electronic noise or distortion in the transmission channel (wireless or wired), and (c) ambient noise in the environment of the listener. Intelligibility is determined by the ratio of the target speech signal to this overall noise level.

The effects of SNR on speech perception are well documented, and, in many cases, it is possible to use the Articulation Index (Kryter, 1962) or the Speech Transmission Index (Steeneken & Houtgast, 1980) to make a quantitative prediction of speech intelligibility directly from the acoustic properties of the noise and speech signals. In general, the sensitivity of speech intelligibility to the SNR depends on the phonetic structure, vocabulary size, and context of the speech signal. Although the CRM phrases provide no contextual information (it is impossible to predict the color or number in a CRM phrase from any of the other words in the phrase), they are limited to a small vocabulary of colors and numbers. This allows listeners to perform well in the CRM task even at very low SNRs. Figure 1

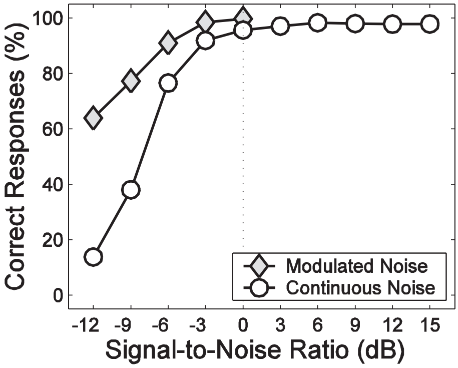

FIGURE 1 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by a continuous or modulated speech-shaped noise signal. Adapted from Brungart (2001).

(adapted from Brungart, 2001) shows performance in the CRM as a function of SNR (calculated for each stimulus as the ratio of the root mean squared (RMS) level measured across the entire individual speech utterance in the stimulus to the long-term RMS level of the individual noise sample in the stimulus) for a continuous speech-shaped noise (circles) and for a speech-shaped noise that has been modulated to match the envelope of a speech signal from the CRM corpus (diamonds). In each case, both the target speech and the noise were presented diotically, that is, with the same audio signal presented simultaneously to both ears. The results show that performance in the CRM task is nearly perfect in continuous noise when the SNR is 0 dB or higher and that performance with a noise masker that is modulated to match the amplitude variations that occur in speech is reasonably good (> 80%) even at an SNR of –6 dB. It should be noted, however, that these surprisingly good results are a direct result of the small vocabulary size in the CRM corpus—the most demanding speech materials (nonsense syllables) require an SNR of approximately +20 dB in the speech band (200 Hz to 6100 Hz) to achieve ≥99% performance (Kryter, 1962). Thus, an ideal multitalker speech display should be able to achieve an SNR of +20 dB in the frequency range from 200 Hz to 6100 Hz (measured from the overall RMS levels of the speech and noise signals). It should be noted that the relative importance of each frequency range to speech intelligibility has been thoroughly documented in the literature on articulation theory (American National Standards Institute, 1969). This information is invaluable when trade-offs between bandwidth and SNR become necessary in the design of communications systems.

Number of Competing Talkers

One obvious factor that can affect the performance of a multitalker speech display is the number of competing talkers. As a general rule, performance in a multitalker listening task decreases when the number of talkers increases. Figure 2 (adapted from Brungart, Simpson, Ericson, & Scott, 2001) shows how performance in the CRM task changes as the number of interfering talkers increases from 0 to 3. The data are shown for different same-sex talkers presented at the same level diotically over headphones. When no competing talkers were present in the stimulus, performance was near 100%. The first competing talker reduced performance by a factor of approximately 0.4 to 62% correct responses. The second competing talker reduced performance by another factor of 0.4 to 38% correct responses, and the third competing talker reduced performance by another factor of 0.4 to 24% correct responses. Thus, one see that CRM performance in a diotic multitalker speech display decreases by approximately 40% for each additional same-sex talker added to the stimulus.

These results clearly show that it is advantageous to reduce the number of simultaneous talkers in a multitalker speech display whenever it is practical to do so.

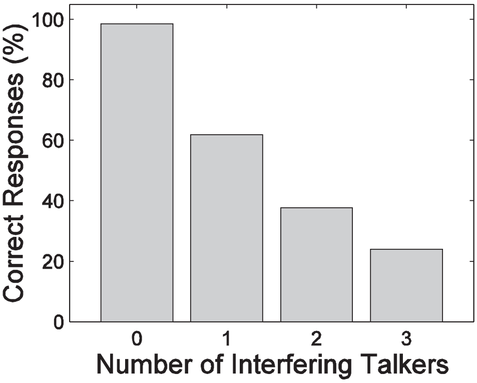

FIGURE 2 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by zero, one, two, or three simultaneous same-sex masking phrases. All of the competing talkers were presented diotically at the same level. Adapted from Brungart, Simpson, Ericson, and Scott (2001).

Possible ways to achieve this reduction range from simple protocols that reduce the chances of overlapping speech signals on a radio channel (such as marking the end of each transmission with a terminator such as “over”), to systems that allow only one talker to speak on a radio channel at any given time, to sophisticated systems that queue incoming messages that overlap in time and play them back to the listener sequentially. However, none of these solutions is appropriate for complex listening situations in which a single communication channel is in near-constant use by two or more simultaneous talkers or situations in which a listener has to monitor two or more communications channels for time-critical information that might occur on any channel. For these situations, the designers of speech displays must rely on other cues to help users segregate the competing speech messages.

Voice Characteristics

Differences in voice characteristics provide one audio cue that can be used to segregate competing speech signals. The voices of different talkers can vary in a wide variety of ways, including differences in fundamental frequency (F0), formant frequencies, speaking rate, accent, and intonation. Talkers who are different in sex are particularly easy to distinguish because on average female talkers have F0 frequencies about two times higher and substantially shorter vocal tracts than male talkers. The shorter vocal tracts of female talkers cause their format center frequencies to be approximately 1.3 times higher than those of male talkers.

Figure 3 (adapted from Brungart et al., 2001) illustrates the effect that differences in voice characteristics can have on a listener’s ability to segregate a target speech signal from one, two, or three interfering talkers. The target and masker

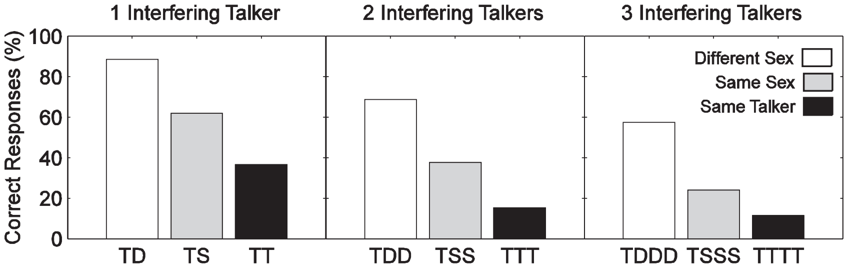

FIGURE 3 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by one, two, or three simultaneous masking phrases. The white bars show performance with masking talkers who were different in sex than the target talker (the TD condition). The gray bars show performance with different masking talkers who were the same sex as the target talker (the TS condition). The black bars show performance when the target and masking phrases were all spoken by the same talker (the TT condition). All of the competing talkers were presented diotically at the same level. Adapted from Brungart, Simpson, Ericson, and Scott (2001).

talkers were randomly selected from the corpus within each block of trials. Thereby, no information about the target or masker talkers' voice characteristics was provided to the listeners. The white bars show performance when the interfering talkers were different in sex than the target talker. The gray bars show performance when the masking phrases were spoken by different talkers who were the same sex as the target talker. The black bars show performance when the target and masking phrases were all spoken by the same talker. In all cases, performance was best when the interfering talkers were different in sex than the target talker and worst when all the phrases were spoken by the same talker.

In situations in which it is possible to control the voice characteristics of the competing talkers in a multitalker speech display, the characteristics of the competing voices should be made as different as possible. One example of a situation in which this should be relatively easy to accomplish is in the use of computer-generated voice icons in an audio display. Consider, for example, a cockpit display where one voice icon might be used to indicate an engine fire and another might be used to indicate a terrain warning. Because the relative priority of these two warnings can vary with the situation, both of these warnings must be presented to the pilot as soon as they occur. If the two warnings are prerecorded in both male and female voices, the display system can act to ensure that the two warnings are spoken by different-sex talkers. This would make it easier for the pilot to attend to the warning with greater immediate relevance.

In audio displays that are designed to present externally generated voice communications rather than internally generated audio icons, it is much more difficult to control the vocal characteristics of the competing talkers. One possible option is to perform some kind of real-time or near-real-time audio processing on the different competing voice signals to make them more distinct. It may be possible to achieve this result by manipulating the parameters used to reconstruct the voice in communication systems that use low-bandwidth parametric vocoders. For example, the F0s of the two talkers could be manipulated to introduce a difference between the two competing talkers in real time. Assman and Summerfield (1990) showed that a difference in F0 of one sixth of one octave is sufficient to produce a significant improvement in intelligibility. However, this approach also has a major drawback: It may make it substantially more difficult (or impossible) for the listener to use voice characteristics to determine the identity of the talker. Thus, the segregation efficiency that is gained by introducing differences in voice characteristics may be more than offset by the reduction in a listener's ability to correctly identify the target talker. A good rule of thumb might be to restrict the use of voice modification to situations in which speaker identification is not important and avoid the use of voice modification when accurate speaker identification is critical. Note also that care must be taken to ensure that voice characteristics such as formant frequencies are not changed enough to degrade the intelligibility of the speech.

Target-to-Masker Ratio (TMR)

Another factor that has a strong influence on a listener's ability to segregate competing speech signals is the level of the target talker relative to the competing talkers. In general, it is much easier to attend to the louder talker in a multitalker stimulus than to the quieter talker in a multitalker stimulus. This is illustrated in Figure 4, which shows performance as a function of the TMR for one, two, or three interfering talkers. In this context, TMR is the ratio of the overall RMS level of the target talker to the overall RMS level of each of the interfering talkers in the stimulus. Thus, when the TMR is 0 dB, all of the talkers in the stimulus are presented at the same level. The results in Figure 4 show that performance is substantially improved when the target talker is the most intense talker in the stimulus (TMR > 0 dB).

Clearly a substantial improvement in speech intelligibility can be achieved by increasing the level of the target talker relative to the levels of the other talkers in the stimulus. Unfortunately, this also degrades the intelligibility of the other talkers in the stimulus. Because it is usually difficult or impossible for the audio display designer to identify the target talker in the stimulus, there is no way to automatically determine which talker should be amplified relative to the others. One alternative approach is to allow the listener to adjust the relative levels of the talkers and thus increase the level of the talker who is believed to be the most important in the current listening situation (Spieth et al., 1954). This ability is provided by current multichannel radio systems, which typically have adjustable level knobs for each

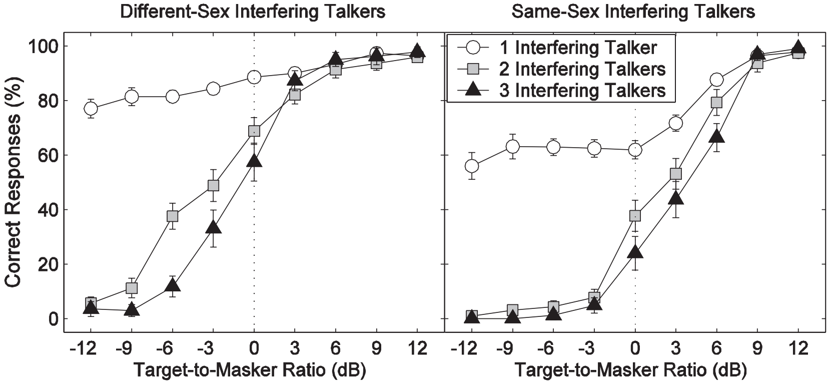

FIGURE 4 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by one, two, or three interfering talkers. The results are shown as a function of the target-to-masker ratio, which is the ratio of the level of the target talker to the level of each of the other interfering talkers in the stimulus (note that all the interfering talkers were presented at the same level). The left panel shows performance with different-sex interfering talkers; the right panel shows performance with same-sex interfering talkers. The error bars show the 95% confidence intervals of each data point. Adapted from Brungart, Simpson, Ericson, and Scott (2001).

radio channel. It should be noted, however, that a potential drawback of this approach is that the listener will miss crucial information that is spoken by one of the low-level talkers in the stimulus: The data in Figure 4 show that performance decreases rapidly with TMR when there are two or more interfering talkers and that listeners essentially receive no semantic information from the low-level talkers when the TMR falls below –6 dB or below 0 dB for same-sex talkers.

The data for the situation with one same-sex interfering talker (open circles in the right panel of Figure 4) have some interesting implications for the design of two-channel communications systems. In this condition, listeners were apparently able to selectively attend to the quieter talker in the stimulus. Consequently, performance in this condition did not decline when the TMR was reduced below 0 dB. Performance did, however, improve rapidly when the TMR was increased above 0 dB. Although one might intuitively expect that two equally important communications channels should be presented at the same level, the data in Figure 4 (adapted from Brungart et al., 2001) suggest that this is a poor strategy. When a level difference is introduced between the two channels, performance improves substantially when the target talker occurs on the louder channel but is unaffected when the target talker occurs on the quieter channel. Thus, overall performance in the CRM task improves substantially when the speech stimuli are presented at levels that differ by 3 dB to 9 dB. These data are also consistent with results of a previous experiment (Egan, Carterette, & Thwing, 1954) that examined performance as a function of TMR with a different call-sign-based task. Note that this strategy only appears to improve performance with same-sex or same-talker interfering speech signals and that it provides much less benefit with different-sex interfering talkers in which differences in voice characteristics seem to dominate any segregation cues provided by differences in the levels of the two talkers. The introduction of level differences may also fail to improve intelligibility in noisy environments where the less intense talker may be masked by ambient noise. Level differences should also be avoided in cases in which there is more than one interfering talker, and intelligibility falls off rapidly with decreasing TMR (Figure 4). Further investigation is needed to explore these level-difference segregation cues in more detail.

BINAURAL FACTORS THAT INFLUENCE MULTITALKER LISTENING

To this point, our discussion has been restricted to factors that influence the performance of monaural or diotic speech displays. When it is possible to use stereo headphones to present a binaural audio signal to the listener, substantial performance benefits can be achieved by using a virtual audio display to spatially separate the apparent locations of the competing sounds (Abouchacra et al., 1997; Crispien & Ehrenberg, 1995; Ericson & McKinley, 1997; Nelson, Bolia, Ericson, & McKinley, 1999). These benefits may be even greater when the listener is provided with some a priori information about the location or voice characteristics of the target talker. In this section, we present the results of a new experiment that examined the effects of spatial separation and a priori information on performance in the CRM multitalker listening task. Because these data have not been presented previously, we present a detailed description of the methods used to collect them in the following.

Method

Listeners. Seven paid volunteer listeners (4 men and 3 women) participated in the study. Their ages ranged from 20 to 55 years with a mean age of 31 years, and all had normal hearing thresholds, that is, less than 20 dB HL from 125 Hz to 8 kHz. Four of these 7 listeners were also participants in the experiments described in the first section of this article (AUTHORS, YEAR).

Apparatus. The stimuli used in this experiment were taken directly from the CRM corpus (Bolia et al., 2000). The speech files were stored on the hard disk of a Pentium-based PC and transferred to a Tucker–Davis Technology array processing card (Tucker-Davis Technologies, Alachua, FL) for playback. A Tucker–Davis Technology Power-SDAC convolved the speech signals with generic head-related transfer functions (HRTFs) using Tucker–Davis Technology's "SOS" HRTFs for the four locations in azimuth without headphone correction. The spatially separated speech phrases were displayed over Sennheiser HD–560. The listeners performed the task while seated in front of the computer monitor in a quiet (≈ 55 dB SPL) room.

Procedure. The data were collected with the same CRM task used in the experiments (Brungart, 2001; Brungart et al., 2001) described in the previous section with three minor variations. The first variation was that only four of the eight call signs in the CRM corpus were used to generate the stimuli in the experiment ("Baron," "Ringo," "Laker," and "Hopper"). The second variation was that the listeners were shown a graphical display with the clock positions of the competing talkers prior to each trial. In some conditions, this display was used to provide information about the location of the target talker. In other conditions, it was simply used to show the locations of all the competing talkers in the stimulus. The third variation was that the listeners responded by pressing labeled keys on a keyboard rather than selecting colored numbers on the screen with a computer mouse.

The test conditions in the experiment were collected in randomly ordered blocks of trials with each block consisting of three consecutive 32-trial sessions with the same level of a priori information. The first session in each block was always conducted with one interfering talker, the second session was always conducted with two interfering talkers, and the third session was always conducted with three interfering talkers. In half of the blocks, male target and masking talkers were used in all three sessions, and in the other half of the blocks, female target and masking talkers were used in all three sessions. Each block of three, 32-trial sessions required an average of 20 min to complete.

The talker locations were fixed at four directions in the horizontal plane. In the one-interfering-talker condition, the two talkers were located at 0° (directly in front of the listener) and –45° azimuth. In the two-interfering-talker condition, potential target and interfering talker locations included 0°, +45°, and –45°. In the three-interfering-talker condition, locations included –45°, 0°, +45°, and +90°. Head tracking was not used in the virtual rendering of the multiple talker display over headphones. Presentation angles of the talkers were therefore fixed with respect to the listener's head.

A total of four different experimental conditions were tested:

- *No a priori information:* In this condition, the listeners were provided with no information about the location or identity of the target talker. They were shown a graphical representation of the clock positions of the talkers used in the stimulus, but they were not told which of these talkers would be the target talker. The target talker and target location were chosen randomly on each trial, and the listener's task was to attend to the phrase containing the call sign "Baron" and respond with the color-number combination contained in that phrase. Each listener participated in eight blocks of three 32-trial sessions in this condition.

- *Known location:* In this condition, the location of the target talker was always fixed at 0°. This location was indicated by a "T" on the graphical display shown to the listeners prior to each trial. The target talker was chosen randomly, and the listener's task was to attend to the talker directly in front (regardless of the call sign used) and respond with the color-number combination contained in that phrase. Each listener participated in eight blocks of three 32-trial sessions in this condition, one with each of the four possible target call signs.
- *Known talker:* In this condition, the same target talker was used throughout each block of trials. This allowed the listeners to learn the identity of the target talker and use this information to help identify the target phrase in each trial. Participants reported being able to learn the target talker's voice characteristics after one or two trials in the one interfering talker condition. The target talker's voice was thereby known for the remaining 30 or 31 trials of the first session and all of Session 2 and 3. The target location was selected randomly in each trial, and no information was given about the location of the target talker. The listener's task was to attend to the target phrase containing the call sign "Baron" and respond with the color-number combination contained in that phrase. Each listener participated in eight blocks of three 32-trial sessions in this condition, consisting of one set with each of the eight possible target talkers.
- *Known talker and known location:* In this condition, the same target talker and target location were used throughout each block of three 32-trial sessions. The location of the target talker (which was selected randomly prior to each block) was indicated by a "T" on the graphical representation of the competing talker locations. The listener's task was to attend to the target phrase containing the call sign "Baron" and respond with the color-number combination contained in that phrase. Each listener participated in eight blocks of three 32-trial sessions in this condition.

Results

Figure 5 shows the effect of spatial separation on overall performance with one, two, or three same-sex interfering talkers. The spatialized results are shown for the condition with no a priori information averaged across all the different target talker locations used in the experiment. The diotic results are taken from a previous experiment (Brungart et al., 2001) that used the same CRM task, the same CRM corpus, and included 4 of the 7 listeners used in this experiment. In the case with one interfering talker, spatial separation increased performance by approximately 25 percentage points. In the cases with two or three interfering talkers, spatial separation nearly doubled the percentage of correct responses. These results clearly illus-

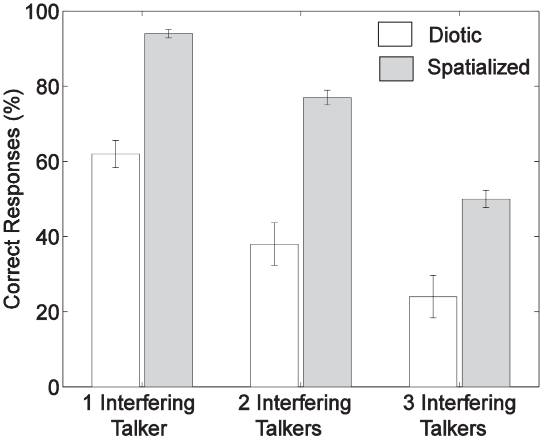

FIGURE 5 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by one, two, or three same-sex interfering talkers. The white bars show results for a diotic condition in which the competing talkers were not spatially separated (adapted from Brungart, Simpson, Ericson, & Scott, 2001). The gray bars show performance in which the competing talkers were spatially separated by 45° (talkers at 0° and 45° with one interfering talker; –45°, 0°, and 45° with two interfering talkers; and –45°, 0°, 45°, and 90° with three interfering talkers). The spatialized results have been averaged across all the different possible target talker locations in each configuration.

trate the substantial performance advantages that spatial separation in azimuth can produce in multitalker audio displays.

Figure 6 shows the effects of a priori information on overall performance with one, two, or three same-sex interfering talkers. The white bars show results from the no a priori data condition in which the target talkers and target locations were chosen randomly in each trial. The gray bars show the results of the known talker condition in which the same target talker was used throughout each block of trials, but the location of the target talker was selected randomly. The black bars show the results of the known talker, known location condition in which the target talker and target location were fixed in each block of trials, and the listeners were shown the location of the target talker prior to each stimulus presentation. The results show that overall performance increased systematically as the listeners were provided with more information about the location and identity of the target talker. When the stimulus contained two or more interfering talkers, overall performance was approximately 20 percentage points higher in the known talker, known location condition than in the condition in which the listeners were provided with no a priori information.

Figure 7 shows a more detailed analysis of the experiment with separate results for each level of a priori information and each possible target talker location. The

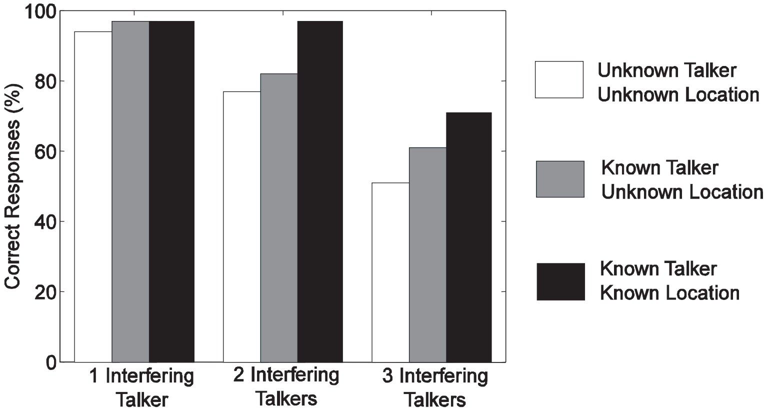

FIGURE 6 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by one, two, or three same-sex interfering talkers. In each case, the talkers were spatially separated by 45° (same configurations as in Figure 5). The white bars show results for a condition in which the listeners had no information about the location or identity of the target talker. The gray bars show a condition in which they knew who the target talker was but not his or her location. The black bars show a condition in which both the identity and location of the target talker were known in advance.

top panel shows performance with one interfering talker, the middle panel shows performance with two interfering talkers, and the bottom panel shows performance with three interfering talkers. For comparison purposes, results from a previous experiment (Brungart et al., 2001) employing the same task for diotically presented stimuli with the same number of interfering talkers are shown. Note that the known-location condition of the experiment was tested only when the target talker was located at 0°.

When the stimulus contained only one interfering talker (Figure 7, top panel), performance was excellent (> 90% correct responses) for all the spatialized talker configurations and a priori information levels tested. Although overall performance was statistically better at 0° than it was at –45° (significant at the $p < .05$ level in a two-factor, within-subjects ANOVA on the arcsine transformed data), the difference between the two conditions was small (< 5%). The level of a priori information had no significant effect on performance. It appears that spatial separation alone improves performance in the one-interfering-talker condition to such a high level that a priori information provides few additional benefits.

When the stimulus contained two interfering talkers (Figure 7, middle panel), the level of a priori information had a much larger impact on overall performance, $F(2, 12) = 32.204$, $p < .001$ in a two-factor, within-subjects ANOVA on the arcsine transformed data). The main effect for location and the interaction of location and a priori information was not significant at the $p = .05$ level. A post hoc analysis of

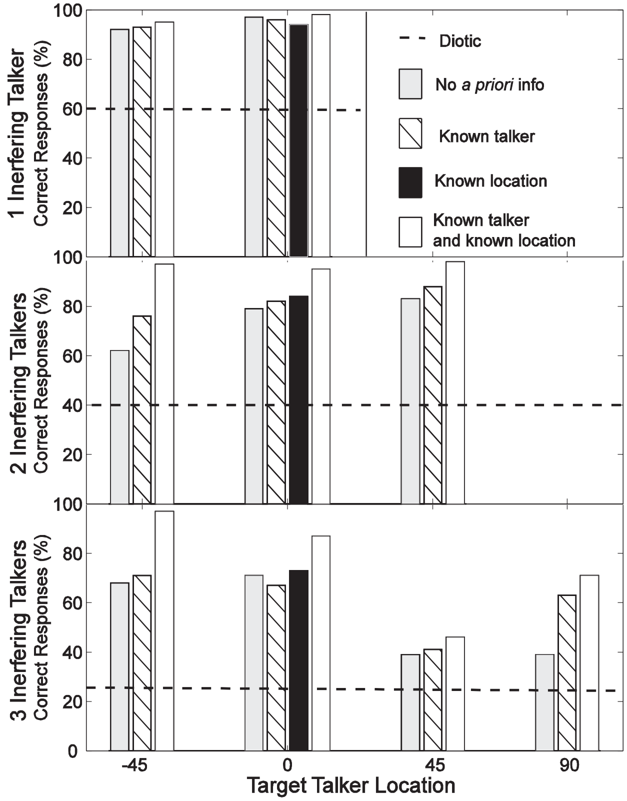

FIGURE 7 Effects of spatial location and a priori information on performance in a multitalker listening task with one competing talker (top panel), two competing talkers (middle panel), and three competing talkers (bottom panel). The bars represent the percentages of correct color and number identifications in trials in which the target talker was located at the indicated target position. For comparison, the results from a previous study (Brungart, Simpson, Ericson, & Scott, 2001) in which the target and masking talkers are presented diotically are also shown (dashed line). Note that the known-location condition was tested only for a target talker at 0°.

the main effect for a priori information found significant differences between no a priori information and known talker at $p = .003$, no a priori information and both cues at $p < .001$, and known talker and both cues at $p < .001$. Performance with a known talker was consistently better than performance with no a priori information, and performance with a known talker and a known location was substantially better than performance with only a known talker (or, at 0°, performance with only a known location.) Although there were some small variations in overall performance across the three talker locations, these variations were not significant at the $p < .05$ level.

When the stimulus contained three interfering talkers (Figure 7, bottom panel), performance varied substantially across the different levels of a priori information and different target talker locations in the experiment. A priori information provided relatively little benefit when the target talker was at +45°, and the target talker was difficult to hear even when its location was known. A priori knowledge of the target talker provided a relatively large benefit when the target talker was at 90°. There was a significant interaction between a priori information and target location, $F(6, 36) = 6.262$, $p < .001$. Performance improved systematically as the level of a priori information increased. Performance also varied systematically with the location of the target talker. Performance was best when the target talker was at –45° and was worst when the target talker was at 45°.

Although these results are complicated, they do provide some valuable insights into the design of improved multitalker speech displays. One important aspect of the results is the universal benefit that was provided by spatial separation of the talkers. Even when the target talker was located at the worst possible spatial location (+45° in the configurations with three interfering talkers), performance in the spatialized condition was still substantially better than in the corresponding diotic condition with the same number of talkers. This is an important advantage of spatial separation over most other methods for improving the intelligibility of a target talker: Most of the audio display strategies that improve the intelligibility of one talker in a multitalker stimulus do so at the expense of one or more of the other talkers in the stimulus. For example, the performance versus TMR curves shown in Figure 4 show that the strategy of presenting critical verbal alarms at least 20 dB above the speech interference level suggested in section 5.3.5.2 of MIL–STD–1472E (U.S. Department of Defense, 1998) can improve the intelligibility of the more intense talker but only at the cost of a decrease in the intelligibility of other talkers in the stimulus. The strategy of low-, high-, and all-pass filtering three competing speech signals suggested in section 5.3.8.2.2 of MIL–STD–1472E increases the intelligibility of the all-pass filtered talker but may eliminate important spectral information from the other two talkers. In contrast, spatial separation improves performance for all channels of the system at the same time.

The results also suggest that performance could be improved by selecting a different set of locations for the talkers in the four-talker condition. When the target talker in the four-talker condition was at 0°, performance was nearly as good as in the three-talker condition. When the target talker was at +45°, however, performance was much worse than in the three-talker condition. This suggests that performance could have been improved by moving the talker located at 45° closer to the talker located at 0° in the four-talker condition. This is consistent with the localization cues listeners use to determine the azimuth locations of sounds: Listeners are much more sensitive to changes in the angle of a sound source near 0° azimuth than to changes in the angle of a sound source near 90° (Mills, 1958). Thus, it could be argued that the 45° talker was "perceptually" much closer to the 90° talker than to the 0° talker and that performance could be improved by moving the 45° talker to a position that was closer to the perceptual midpoint between these two locations (perhaps 30°).

A final consistent finding in these results is that a priori knowledge about both the talker and location of the target phrase provides a much larger performance benefit than a priori information about just the target talker or a priori information about just the target location. A post hoc, least significant difference analysis was performed on the main effect of a priori knowledge. The combination of a priori talker and location information was different than the no a priori information condition ($p < .001$), a priori talker condition ($p < .001$), and the a priori location information ($p < .001$). No other paired comparisons produced statistically significant differences. This suggests that audio display designers should strive to ensure that the different talkers in a multitalker speech display are consistently placed in the same locations so that a listener who chooses to attend to one particular channel for a long period of time will know for whom and from where to listen. Beyond this, however, it is probably not practical to make use of the advantages of a priori information in multitalker audio displays. If the system somehow knew which talker the listener would be listening to, the optimal strategy would be to eliminate the other talkers from the stimulus. The primary reason for having multichannel speech displays is that neither the user nor the system designer knows which channel will provide the most useful information at any given time: All channels must be monitored at all times to ensure that important information is not lost. Thus, is it unlikely that a display will be able to reliably indicate which channel to listen to in any given situation.

Care should, however, be taken in the evaluation of multitalker speech displays to accurately model the amount of a priori information that will be available to the eventual end user of the system. In some operational tasks, the target talker will change frequently, and the listener must always monitor all the channels vigilantly. In other tasks, the listener will engage in a conversation with a single talker for a long period of time before switching attention to one of the other channels of the system. Failure to account for the differences in these two situations may prevent an accurate assessment of the true operational effectiveness of the system.

EFFECTS OF SPATIAL SEPARATION IN NOISY ENVIRONMENTS

The advantages of spatial separation can be even more pronounced in a noisy environment. Figure 8 (adapted from Ericson & McKinley, 1997) shows the effect of spatial separation with one same-sex competing talker as a function of the amount

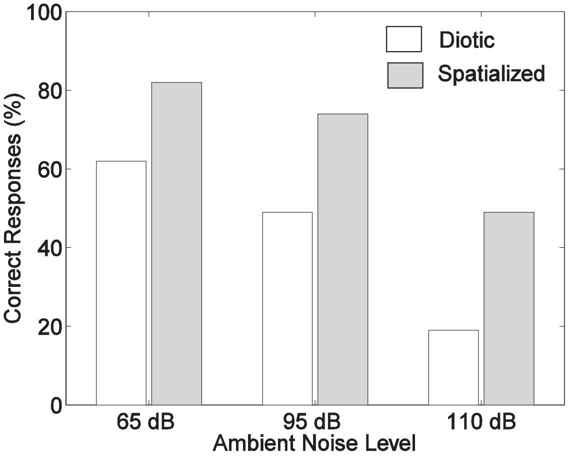

FIGURE 8 Percentage of correct color and number identifications for a Coordinate Response Measure target phrase masked by one same-sex interfering talker. The white bars show results for a diotic condition in which the competing talkers were not spatially separated. The gray bars show performance when the competing talkers were spatially separated by 45°. Note that the target and masking phrases were spoken by live talkers wearing an aviation helmet and oxygen mask and that these signals were passed through a military intercom system before being spatially processed and presented to the listeners over Bose AH–A (Bose Corp., Framingham, MA) active noise reduction headsets. Adapted from Ericson and McKinley (1997).

of ambient noise in the environment. These data are taken from an experiment in which the CRM phrases were spoken by live talkers wearing oxygen masks and heard by listeners wearing active noise reduction headsets with both the talkers and listeners immersed in the ambient noise. The results show that the advantages of spatial separation were greatest when the listeners were subjected to an ambient noise field of 110 dB SPL (84 dB SPL under the earcup). This should be taken into consideration in the design of displays for use in noisy environments.

EFFECTS OF SPATIAL SEPARATION IN DISTANCE

The advantages of spatial separation are not limited to direction. Recent experiments (e.g., Brungart & Simpson, 2001) have shown that substantial improvements in performance can also be achieved by spatially separating nearby talkers in distance in the near-field region where listeners can use binaural cues to determine the distances of sound sources. In Brungart and Simpson's experiment that used phrases from the CRM corpus that were spatially processed with near-field HRTFs measured at 90° in azimuth, the percentage of correct color and number identifications increased from 55% when the target and masking talkers were presented at the same distance (1 m) to more than 80% when one talker was presented at 1 m and the second talker was presented at 12 cm.

CONCLUSIONS

The most efficient way to improve the effectiveness of a multitalker speech display is to use virtual synthesis techniques to spatially separate the locations of the competing talkers. The data from Figure 5 show that spatially separating same-sex competing talkers by 45° produced a 25 to 35 percentage point increase in overall performance in the CRM task. In terms of the other factors examined in this article, this is roughly equivalent to (a) reducing the number of competing talkers in the stimulus by 1 to 1.5 talkers (Figure 2), (b) replacing the same-sex interfering talkers with different-sex interfering talkers (Figure 3), or (c) increasing the TMR by 3 dB to 9 dB (Figure 4).

However, spatial separation has substantial advantages over these other techniques. The biggest advantage is that spatial separation improves the intelligibility of all the talkers in the stimulus, whereas the other techniques tend to increase the intelligibility of only one of a few selected talkers. Reducing the number of talkers in the stimulus increases the intelligibility of the remaining talkers at the expense of losing all the information from the eliminated talker. Replacing the same-sex interfering talkers with different-sex talkers provides a benefit only for the talker who is different in sex from the other talkers in the stimulus. Increasing the TMR increases the intelligibility of one talker but generally reduces the intelligibility of the other talkers in the stimulus when there are more than two talkers. Only spatial separation is able to improve overall performance across all the talkers in a three- to four-talker stimulus.

Spatial separation is also relatively inexpensive to implement in multitalker speech displays. Many of the benefits of spatially separating speech signals can be obtained with relatively simple digital signal processing techniques that do little more than introduce interaural time differences (Carhart, Tillman, & Johnson, 1967) and interaural level differences (Bronkhorst & Plomp, 1988) into the different communications channels of the system. The listener-specific, pinna-related spectral details that are required to produce realistic, localizable, externalized virtual sounds in nonspeech virtual displays (Wenzel, Arruda, Kistler, & Wightman, 1993) simply do not provide any additional benefit to speech intelligibility in multitalker listening tasks for presentation in azimuth (Drullman & Bronkhorst, 2000; Nelson et al., 1999). Similarly, real-time head-tracking devices are not required to achieve good intelligibility in multitalker speech displays (the data shown in Figure 5 were collected without any head tracking). If a communications system or intercom is capable of processing audio signals in the digital domain, it may be possible to implement an effective speech segregation algorithm in software for little or no additional cost. The only restriction is that the system must be capable of producing a stereo output signal: No reliable spatial cues are possible in a system with only one analog output channel.

If the entire audio system is accessible to the audio display designer, spatial separation is clearly the best technique for improving the performance of the system. Unfortunately, many existing systems have architectural constraints that prevent the installation of a binaural speech display. Under these conditions, alternative methods must be used to improve the performance of the audio display system. Clearly every effort should be made to ensure that only time-critical speech signals are allowed to share the communication channel at the same time; noncritical signals should be eliminated or delayed until the channel is open. Real-time modification of the voice characteristics of the talkers may provide some performance benefits, but this technology is far from mature at this time. Because no single strategy will work for all single-channel speech displays, audio display designers must carefully consider the specific tasks the listener will perform with the system and try to tailor the display for those tasks.

The availability of a priori knowledge about the target talker's voice and his or her location can significantly improve speech intelligibility in spatially separated, multitalker listening conditions. This improvement was greater with two and three interfering talkers than with one interfering talker. Listeners appeared to be able to divide their attention between two simultaneous talkers and efficiently monitor both talkers for the target call sign, but they had difficulty monitoring three or four simultaneous talkers without some a priori knowledge about the target talker's voice characteristics or location. A priori knowledge of the target talker's voice and location was better than knowing either the voice characteristics or location alone. Although it is difficult to take direct advantage of a priori information in the design of multitalker speech displays, these results suggest that intelligibility may be better when fixed locations are assigned to each of the competing channels in the system than when channels change location dynamically. Note, however, that some displays may use the locations of the speech signals to convey spatial information to the listener and that the benefits of this information may outweigh the costs of a small loss in speech intelligibility.

Although in this article we have reviewed many of the factors that can influence the performance of a multitalker speech display, we have by no means explored all of these issues. Further investigation is needed to determine how the different display techniques outlined in this article interact with one another. More research is needed to determine the optimal locations of the talkers in a spatialized speech display: Most researchers have placed the competing talkers at evenly spaced locations in azimuth, but no systematic studies have been conducted to determine if this placement is ideal. Other factors, such as the effect of talker motion on speech segregation or the benefits that can be obtained by adding real-time head tracking to a multitalker speech display, also require further exploration. Finally, greater efforts must be made to determine how multitalker displays can be tailored for the specific communication tasks they are designed to address.

Communications tasks can vary widely in terms of vocabulary size, speech syntax, and available contextual information. Communication tasks can also vary in terms of how frequently the listener is required to switch attention across the different competing talkers and in terms of the nonspeech tasks listeners are required to perform concurrently with the communication task (Tun & Wingfield, 1994). At this point, most research in multitalker speech displays has been focused on "general-purpose" communications tasks. New techniques are needed to develop and test speech displays for more specific applications. Only when these issues are resolved will it be possible to begin converging on a series of protocols for designing truly optimal multitalker speech displays.

ACKNOWLEDGMENTS

We thank Richard McKinley for his helpful comments on this manuscript.

REFERENCES

Abouchacra, K., Tran, T., Besing, J., & Koehnke, J. (1997, February). 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.

American National Standards Institute. (1969). A*NSI 53.5–1969: Methods for calculation of the articulation index.* New York: Acoustical Society of America.

Assman, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America, 88,* 680–697.

Begault, D. R. (1999). Virtual acoustic displays for teleconferencing: Intelligibility advantage for "telephone-grade" audio. *Journal of the Audio Engineering Society, 47,* 824–828.

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

Bronkhorst, A., & Plomp, R. (1988). The effect of head-induced interaural time and level difference on speech intelligibility in noise. *Journal of the Acoustical Society of America, 83,* 1508–1516.

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

Brungart, D., & Simpson, B. (2001). Optimizing multitalker speech displays with near-field HRTFs. In J. Hiipakka, N. Zacharov, & T. Takala (Eds.), *Proceedings of the 2001 International Conference on Auditory Display* (pp. 169–174). Espoo, Finland: Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory.

Brungart, D., Simpson, B., Ericson, M., & Scott, K. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America, 110,* 2527–2538.

Carhart, R., Tillman, T., & Johnson, K. (1967). Release from masking for speech through interaural time delay. *Journal of the Acoustical Society of America, 42,* 124–138.

Crispien, K., & Ehrenberg, T. (1995). Evaluation of the “cocktail party effect” for multiple speech stimuli within a spatial audio display. *Journal of the Audio Engineering Society, 43,* 932–940.

Drullman, R., & Bronkhorst, A. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America, 107,* 2224–2235.

Egan, J., Carterette, E., & Thwing, E. (1954). Factors affecting multi-channel listening. *Journal of the Acoustical Society of America, 26,* 774–782.

Ericson, M., & McKinley, R. (1997). The intelligibility of multiple talkers spatially separated in noise. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 701–724). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kryter, K. (1962). Methods for calculation and use of the Articulation Index. *Journal of the Acoustical Society of America, 34,* 1689–1697.

Mills, A. (1958). On the minimum audible angle. *Journal of the Acoustical Society of America, 30,* 237–246.

Moore, T. (1981). Voice communication jamming research. In *AGARD Conference Proceedings 331: Aural communication in aviation* (pp. 2:1–2:6). Neuilly-Sur-Seine, France: PUBLISHER NAME.

Nelson, W. T., Bolia, R. S., Ericson, M. A., & 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). LOCATION: PUBLISHER NAME

Spieth, W., Curtis, J., & Webster, J. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustical Society of America, 26,* 391–396.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America, 67,* 318–326.

Tun, P., & Wingfield, A. (1994). Speech recall under heavy load conditions: Age, predictability and limits on dual-task interference. *Aging, Neuroscience, and Cognition, 1,* 29–44.

U.S. Department of Defense. (1998). MIL–STD–1472E, Department of Defense Design Criteria Standard. Washington, DC: Government Printing Office.

Wenzel, E., Arruda, M., Kistler, D., & Wightman, F. (1993). Localization using non-individualized head-related transfer functions. *Journal of the Acoustical Society of America, 94,* 111–123.

Manuscript first received June 2003