Improving Multitalker Speech Communication with Advanced Audio Displays

Douglas S. Brungart and Brian D. Simpson
Air Force Research Laboratory, AFRL/HECB, Bldg 441, 2610 Seventh Street,
Wright-Patterson AFB, OH 45433-7901 Tel: (937) 255-3660 ext 422 Fax: (937) 656-7680

douglas.brungart@wpafb.af.mil

ABSTRACT

Historically, most of the metrics that have been used to evaluate the effectiveness of military communications systems have focused on measuring the intelligibility of a single talker in the presence of a continuous noise masker. However, many critical military operations involve complex communications tasks that require listeners to monitor, process, and respond to two or more simultaneous speech signals. Many factors can influence performance in such tasks, including the relative levels of the competing talkers, the similarities between the voices of the competing talkers, and, in audio displays that allow the different channels of speech to be spatially separated, the apparent locations of the target and masking voices. In this paper, we present an overview of the factors that can influence speech intelligibility in multitalker listening environments, and compare and contrast them to the factors that influence intelligibility in the speech-in-noise situations that are usually used to evaluate military communications systems. We also discuss the intelligibility benefits that can be achieved with advanced audio displays that use either dichotic or binaural processing to spatially separate the apparent locations of multiple simultaneous channels of speech. Finally, we describe a spatial configuration that has been shown to maximize the benefit of spatial separation in a listening configuration with up to seven simultaneous speech signals.

1.0 INTRODUCTION

Many critically important military 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 military air traffic control centers, where controllers are 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. In command and control centers, commanders are also faced with the task of assimilating information obtained from multiple sources and using this information to coordinate multi-unit actions against enemy assets. Military pilots face an even more difficult situation where they may need to communicate with other personnel in their own aircraft, other aircraft in their own formation, command and control personnel in AWACS aircraft and at ground-based command centers, and ground-based target spotting personnel near the site of an airstrike. 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 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

Brungart, D.S.; Simpson, B.D. (2005) Improving Multitalker Speech Communication with Advanced Audio Displays. In *New Directions for Improving Audio Effectiveness* (pp. 30-1 – 30-18). Meeting Proceedings RTO-MP-HFM-123, Paper 30. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.aps.

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; MIL-STD-1472E, 1997). More recent efforts have used virtual audio displays to spatially separate the competing speech channels (Crispien & Ehrenberg, 1995; Ericson & McKinley, 1997; Begault, 1999). 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: 1) the signal-to-noise ratio in the communications system; 2) the number of competing talkers; 3) the voice characteristics of the talkers; 4) the relative levels of the talkers; 5) the apparent spatial locations of the talkers; and 6) the listener's *a priori* knowledge about the listening environment. A further complicating issue is the variety of methodologies that has been used to examine these factors; procedural variations often make it difficult to compare the results of different multitalker listening experiments. In this paper, we present the results of a number of experiments that have used the Coordinate Response Measure (CRM) to examine the impact that different audio display design parameters has 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.

2.0 EXPERIMENTAL METHODOLOGY: THE COORDINATE RESPONSE MEASURE

All of the experiments described in this paper were conducted using the Coordinate Response Measure (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 (Moore, 1981). 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, white), and one of eight numbers (1-8). The listener's task is to listen for the target sentence containing the pre-assigned 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 all of the simultaneous phrases in order to successfully extract the information contained in the target phrase (Spieth et al., 1954; Abouchacra, Tran, Besing, & Koehnke, 1997). In this regard, it is similar to many command and control tasks where 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 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 X 4 colors X 8 numbers) spoken by eight different talkers (four male, four female). The experiments described 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. Each of the following sections discusses a different factor that influences speech intelligibility in a multitalker listening environment.

3.0 FACTORS INFLUENCING PERFORMANCE IN MULTITALKER DISPLAYS

3.1 Signal-to-Noise Ratio:

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: 1) ambient noise in the environment of the talker that is picked up by the microphone that records the talker's voice; 2) electronic noise or distortion in the transmission channel (wireless or wired); and 3) ambient noise in the environment of the listener. Intelligibility is determined by the ratio of the target speech signal to this overall noise level.

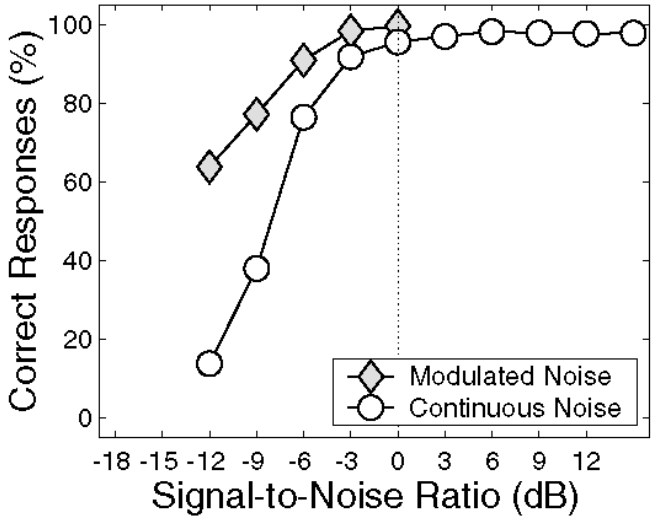

Figure 1: Percentage of correct color and number identifications for a CRM target phrase masked by continuous or modulated speech-shaped noise signal. Adapted from Brungart (2001).

The effects of signal-to-noise ratio (SNR) on speech perception are well documented, and, in many cases, it is possible to use the Articulation Index (AI) or the Speech Transmission Index (STI) to make a quantitative prediction of speech intelligibility directly from the acoustic properties of the noise and speech signals (Kryter, 1962; Steeneken & Houtgast, 1980). 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 (adapted from Brungart, 2001a) shows performance in the CRM as a function of SNR (calculated for each stimulus as the ratio of the 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, i.e., 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 - 6100 Hz) to achieve near-perfect identification 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. This information is invaluable when tradeoffs between bandwidth and SNR become necessary in the design of communications systems.

3.2 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 (Brungart et al., 2001). 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 we 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. 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 like “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 where a single communication channel is in nearly constant use by two or more simultaneous talkers or where 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.

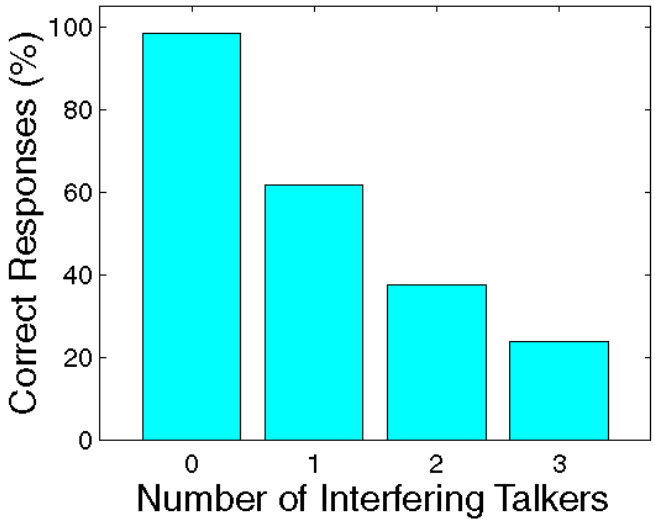

Figure 2: Percentage of correct color and number identifications for a CRM target phrase masked by 0, 1, 2, or 3 simultaneous same-sex masking phrases. All of the competing talkers were presented diotically at the same level. Adapted from Brungart et al. (2001).

3.3 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 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 where 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 where 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

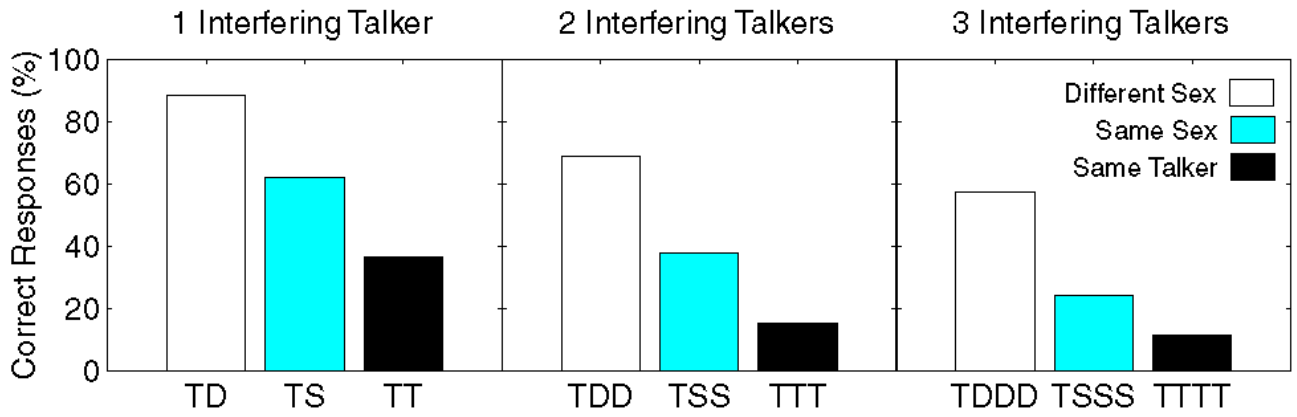

Figure 3: Percentage of correct color and number identifications for a CRM target phrase masked by 1, 2, or 3 simultaneous masking phrases. The white bars show performance with masking talkers who were different in sex than the target talker (the TD condition). The gray bars show performance with different 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 et al. (2001).

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 pre-recorded 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 fundamental frequencies (F0s) of the two talkers could be manipulated to introduce a difference between the two competing talkers in real time. Assman and Summerfield (1990) have shown that a difference in F0 of 1/6th 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.

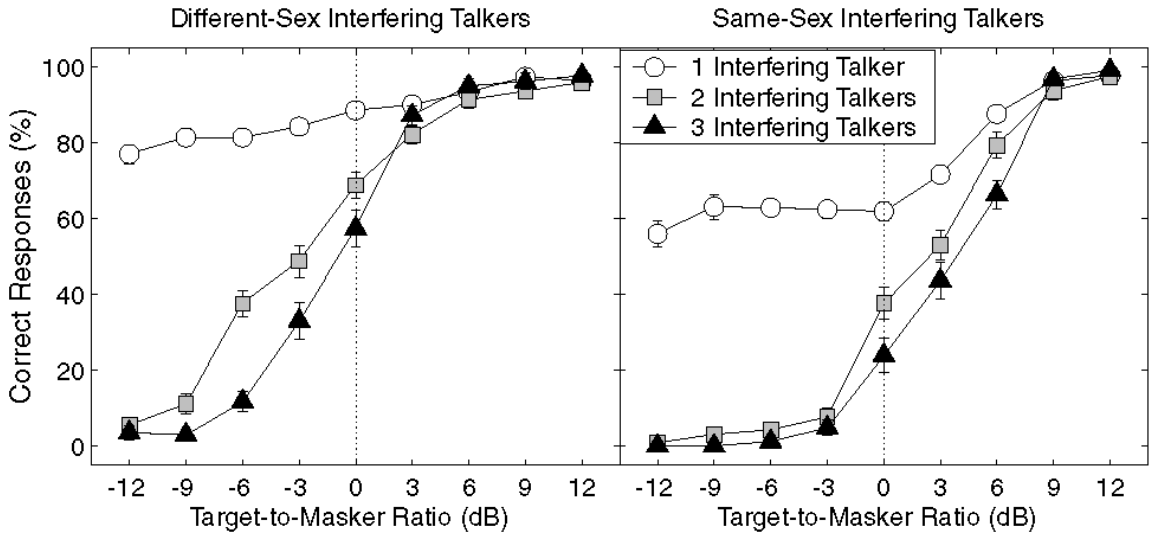

Figure 4: Percentage of correct color and number identifications for a CRM target phrase masked by 1, 2, or 3 interfering talkers. The results are shown as a function of the target-to-masker ratio (TMR), which is the ratio of the level of the target talker to the level of each of the other interfering 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 et al. (2001).

3.4 Target-to-Masker Ratio:

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 target-to-masker ratio (TMR) for 1, 2, or 3 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 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 volume-level knobs for each 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 3-9 dB. These data are also consistent with results of a previous experiment that examined performance as a function of TMR with a different call-sign-based task (Egan, Carterette, & Thwing, 1954). Note that this strategy appears to improve performance only with same-sex or same-talker interfering speech signals, and that it provides much less benefit with different-sex interfering talkers where 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 where 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.

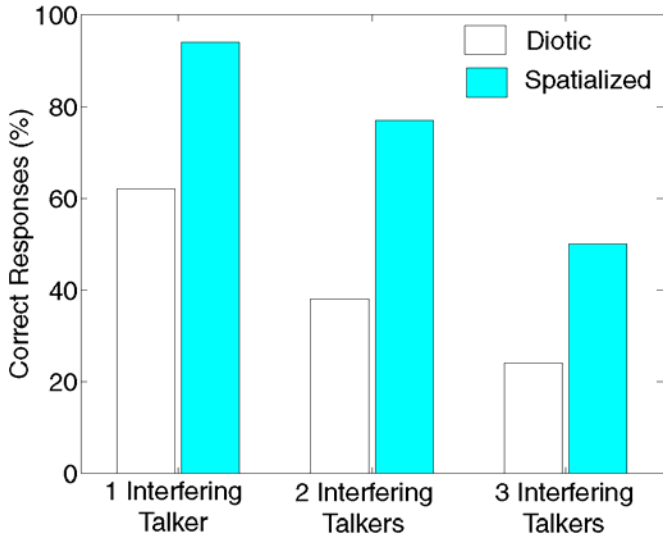

Figure 5: Percentage of correct color and number identifications for a CRM target phrase masked by 1, 2, or 3 same-sex interfering talkers. The white bars show results for a diotic condition where the competing talkers were not spatially separated (adapted from Brungart et al., 2001). The gray bars show performance where 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.

3.5 Spatial Separation:

In intercom systems where 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 (Crispien & Ehrenberg, 1995; Abouchacra et al., 1997; Ericson & McKinley, 1997; Nelson, Bolia, Ericson, & McKinley, 1999). Figure 5 shows the effect of spatial separation on overall performance in the CRM task with one, two, or three same-sex interfering talkers. These data have been taken from an experiment where the target and masking talkers were always separated by 45 degrees (Ericson et al., 2004). 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 illustrate the substantial performance advantages that spatial separation in azimuth can produce in multitalker audio displays.

3.5.1 Target Location

In listening situations where the target and masking talkers are spatially separated, the actual level of intelligibility achieved at any given time will vary according to where the target talker is located and how large the angular separation is between the target talker and the other simultaneous talkers on the intercom. The left panel of Figure 6 shows two possible spatial configurations for a three-channel spatial intercom: a 3-Talker Close configuration, where the competing talkers are separated by 15 degrees, and a 3-Talker Far configuration, where the talkers are separated by 90 degrees. The middle panel of the figure shows mean performance in the CRM task as a function of the location of the target talker (Brungart et al., 2005). Note the following three main points from these results:

- Overall performance is generally better for widely-separated talker locations than it is for closely-spaced talker locations.
- Performance is generally better when the target talker is located at the leftmost or rightmost talker location than when it is located at a central location. This occurs because the target talker is the most intense talker in either the left ear or the right ear when it is located at the extreme leftmost or rightmost position, but is less intense than at least one talker in both ears when it is located at a central location (e.g., see Zurek, 1993).
- Performance is generally better when the target talker is located in the right hemisphere than when it is located in the left hemisphere. This is believed to be related to the specialization of speech processing centers in the left hemisphere of the human brain (Bolia et al., 2001).

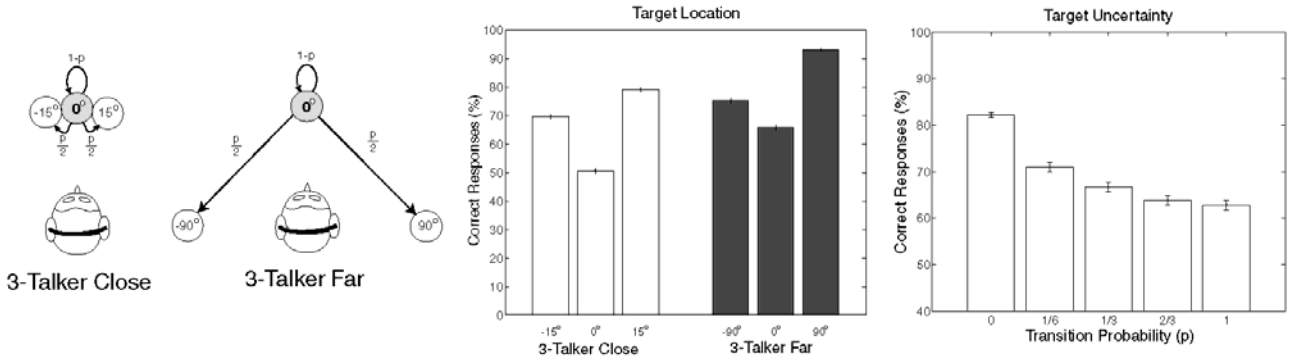

Figure 6: Effects of Spatial Location and Target Uncertainty on Multitalker Listening Peformance. The left panel of the figure illustrates two 3-talker spatial listening environments, one with the competing talkers separated by 15 degrees and one with the competing talkers separated by 90 degrees. The middle panel shows the percentage of correct responses in the CRM task as a function of the target talker location in each of these configurations. The right panel shows performance as a function of the level of uncertainty the listener has about the location of the target talker on any given trial, which is expressed by the probability *p* that the target talker will change from its current location to a different location at the end of any given trial in a block of trials in the experiment (as illustrated in the form of a Markov diagram by the arrows in the left panel of the figure). The data have been adapted from Brungart and Simpson, 2005.

3.5.2 Target Uncertainty

Another factor that influences performance in multitalker communications tasks that involve a spatialized intercom system is the amount of *a priori* information the listener has about where the most important information will originate at any given time. In general, listeners benefit more from the spatial separation of competing channels of speech when they know which talker to listen to than when they have to scan their attention across multiple talkers to determine where the most pertinent information is located. One way to measure the effect this *a priori* information has on performance is to perform a multitalker listening experiment where the likelihood that the target talker will move from its current location to a new location at the end of any given trial is systematically varied across blocks. The arrows in the left panel of Figure 6 illustrate this kind of dynamic listening environment in the form of a Markov diagram where the number associated with each arrow represents the probability that the system state will change in a particular way at the end of a single trial of the experiment. The bars in the right panel of Figure 6 show how performance varies in the 3-talker CRM task as a function of the total transition probability p of the target talker moving from its current location to a new location at the end of each trial. These results show that listeners perform best when the target talker is located in a fixed position (p=0), that their overall performance degrades substantially (by about 12 percentage points in this case) when a small amount of uncertainty is added to their system, and that their performance degrades more gradually as additional uncertainty is added up to the point where the target talker location changes in every trial of the experiment (p=1) (Brungart and Simpson, 2005).

3.5.4 Headtracking

A final factor that has the potential to influence performance in spatialized multitalker speech displays is the integration of a real-time headtracking device that updates the apparent locations of the competing talkers in the system to compensate for the movements of the listener's head. This kind of head-coupling is known to improve the fidelity, realism, and localizability of virtual sound sources (e.g., see Wallach 1940), so it might be reasonable to expect it also to improve performance in multitalker listening tasks. However, recent experiments in our laboratory (Brungart and Simpson, 2004) have shown that real-time headtracking provides, at most, a modest performance benefit in multitalker listening tasks, and that these benefits are limited to those cases where the target talker location changes infrequently (i.e., the transition probability is low) and the competing talkers are widely separated in azimuth. Because headtracking is quite expensive to implement, these results suggest that it is not likely to be cost effective to implement in military communications applications that do not already have access to real-time head position information for some non-communication related purpose (such as a helmet-mounted display in an aircraft cockpit).

4.0 AN OPTIMIZED SPATIAL CONFIGURATION FOR SEVEN TALKERS

4.1 Choosing an Optimal Spatial Configuration for a Multitalker Display:

Table 1: Summary of locations used in previous multitalker speech displays

Study	# of Talkers	Talker Locations
1) Cherry (1953)	2	Non-spatial (left ear only, right ear only)
2) Triesman (1964)	3	Non-spatial (left ear only, right ear only, both ears)
3) Moray et al. (1964)	4	Non-spatial (L only, 2/3 L+1/3 R; 1/3 L + 2/3 R; R only)
4) Khoenke et al. (1998)	4	-90, -45, +45, +90 azimuth
5) Spieth et al. (1954)	3	-20,0,20 azimuth or –90,0,90 azimuth
6) Drullman & Bronkhorst (2000)	4	-90, -45, 0, +45, +90
7) Yost (1996)	7(3)	-90, -60, -30, 0, +30, +60, +90 azimuth
8) Hawley et al. (1999)	7(2-4)	-90, -60, -30, 0, +30, +60, +90 azimuth
9) Crispien & Ehrenberg (1995)	4	-90 az, +60 el; -30 az, +20 el; -30 az, -20 el; -90 az, -60 el
10) Nelson et al. (1999)	8(2-8)	6: -90, -70, -31, +31, +70, +90 7: -90, -69, -45, 0, +45, +69, +90 8: -90, -69, -45, -11, +11, +45, +69, +90 azimuth
11) Simpson et al. (1998)	8(2-8)	7: -90, -69, -135, 0, +135, +69, +90 8: -90, -69, -135, -11, +11, +135, +69, +90 azimuth
12) Ericson & McKinley (1997)	4	-135, -45, +45, +135 azimuth (w/headtracking)
13) Brungart & Simpson (2002)	2	90 degrees azimuth, 1 m; 90 degrees azimuth, 12 cm

Although a number of researchers have demonstrated the advantages of spatial filtering for multitalker speech perception, very little effort has been made to systematically develop an optimal set of HRTF filters capable of maximizing the number of talkers a listener can simultaneously monitor while minimizing the amount of interference between the different competing talkers in the system. Most systems that have used HRTF filters to spatially separate speech channels have placed the competing channels at roughly equally-spaced intervals in azimuth in the listener's frontal plane. Table 1 provides examples of the spatial separations used in previous multitalker speech displays. The first three entries in the table represent early systems that used stereo panning over headphones rather than head-related transfer functions to spatially separate the signals. This method has been shown to be very effective for the segregation of two talkers (where the talkers are presented to the left and right earphone), somewhat effective for the segregation of three

talkers (where one talker is presented to the left ear, one talker is presented to the right ear, and one talker is presented to both ears), and only moderately effective in the segregation of four talkers (where two talkers are presented to the left and right ears, one talker is presented more loudly in the left ear than in the right ear, and one talker is presented more loudly in the right ear than the left ear). However, these panning methods have not been shown to be effective in multitalker listening configurations with more than four talkers.

The other entries in the table represent more recent implementations that either used loudspeakers to spatially separate the competing speech signals or used HRTFs that accurately reproduced the interaural time and intensity difference cues that occur when real sound sources are spatially separated around the listener's head. The majority of these implementations (entries 4-8 in Table 1) have used talker locations that were equally spaced in the azimuth across the listener's frontal plane. One implementation (entry 9 in Table 1) has spatially separated the speech signals in elevation as well as azimuth, varying from +60 degrees elevation to –60 degrees elevation as the source location moves from left to right. And two implementations (entries 10 and 11 in Table 1) have used a location selection mechanism proposed by Nelson et al. (1999) that selects talker locations in a procedure designed to maximize the difference in source midline distance (SML) between the different talkers in the stimulus.

Recently, a new talker configuration has been proposed in which the target and masking talkers are located at different distances (12 cm and 1 m) at the same angle in azimuth (90 degrees) (entry 13 in Table 1). This spatial configuration has been shown to work well in situations with only two interfering talkers, but it has not been tested with more than two competing talkers.

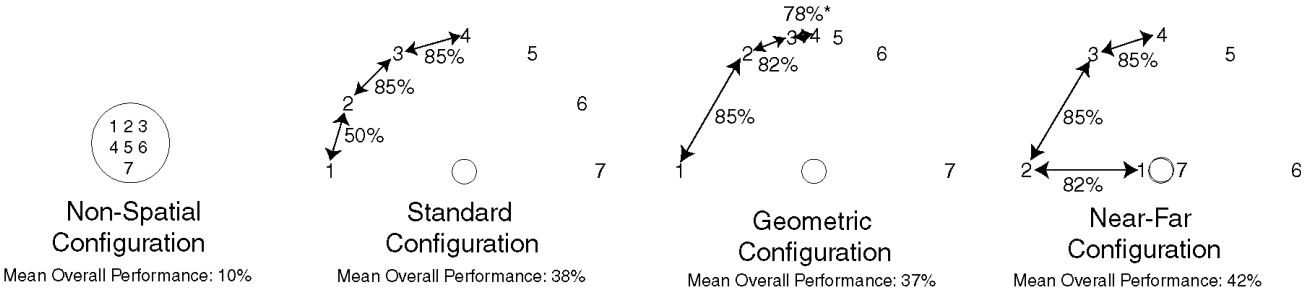

Figure 6: Comparison of performance for different spatial configurations in a multitalker speech display system with seven simultaneous talkers. The leftmost panel shows a standard monaural configuration where all of the talkers are presented diotically and appear to originate from the center of the listener's head. The next panel shows a standard spatial configuration where the talkers are equally spaced 30 degrees apart in azimuth. The third panel shows a geometric configuration where the talkers are spaced closely together in front of the listener (where spatial resolution is greatest) and further apart to the sides (where spatial resolution is worst). The right panel shows a hybrid near-far configuration with five geometrically-spaced far-field talkers plus one nearby talker in each of the listener's ears. The arrows show mean performance in a two-talker task CRM with competing talkers at each end of the arrows, and the overall performance numbers at the bottom show performance in a CRM task with seven simultaneous same-sex talkers (adapted from Brungart and Simpson, 2003).

While it is possible to make theoretical arguments in favor of each of these possible talker configurations, we know of no previous studies that have objectively measured speech intelligibility as a function of the placement of the competing talkers. However, recent results in our laboratory have shown that equal spacing in azimuth cannot produce optimal performance in systems with more than five possible talker locations. The reason for this stems from the fact that listeners are generally more sensitive to changes in the spatial locations of objects in front of them than they are to changes in the spatial locations of objects located off to the sides (Mills, 1950). Thus it turns out that, in a seven-talker configuration with the talkers spread out equally every 30 degrees in the listener's front hemifield (the "standard" configuration in Figure 6), listeners are generally good at segregating speech signals that happen to occur from adjacent talkers near 0 degrees azimuth [as indicated by the 85% correct performance level achieved in a two-talker listening test with talkers at locations 3 and 4 (indicated by the arrow in the figure)], but they are very poor at segregating adjacent competing talkers at lateral locations (as indicated by the 50% performance level for the 2-talker CRM task with talkers at locations 1 and 2). This inadequate spatial resolution at lateral source positions suggests that better overall performance might be achieved with a "geometrically" spaced source configuration with closely-spaced talker locations near the midline and larger angular separations between the talkers at more lateral locations (third panel of Figure 6). Such a configuration does indeed result in more uniform (and better) segregation performance for pairs of adjacent talker locations, but unfortunately it does not result in improved performance when all seven competing talkers are active at the same time (indicated by the overall performance scores underneath each panel in Figure 6).

Significantly better overall performance *can,* however, be achieved with the seven-talker hybrid near-far configuration shown in the rightmost panel of Figure 6. This configuration takes advantage of the fact that listeners can use interaural level differences to distinguish between nearby and distant talkers located along the interaural axis (Brungart and Simpson, 2002), and provides a configuration that produces both a high level of performance for pairs of adjacent talkers in a two-talker CRM task and a roughly 10% improvement in overall performance (relative to the standard configuration) in a seven-talker CRM task. Our research indicates that this configuration represents a near-optimal configuration for a multitalker display with more than five simultaneous talkers.

4.2 The Benefits of Spatial Audio in a Realistic Military Communications Environment:

While there is general consensus that spatialized audio can improve multitalker listening performance in laboratory settings, it is not necessarily always clear that these benefits will transfer in a useful way to operational military environments, where the communications situation is dynamically changing and the number of simultaneous talkers varies across time. Also, while it is indisputably true that spatialized audio can improve multitalker intelligibility relative to a monaural speech display that provides the same signal to both of the listener's ears, it is not at all obvious that spatialized displays will provide a compelling benefit over *dichotic* speech displays that provide the listener with the option of hearing competing radio channels in the left ear only, the right ear only, or in both ears simultaneously (which effect causes them to be perceived in the "center of the head"). In order to test these situations, we conducted an experiment in our laboratory that compared the performance of monaural, dichotic, and spatialized speech displays in a seven-talker CRM listening task where each talker was 50% likely to be active on any given trial and the target phrase had a 25% probability of changing to a different talker at a different location at the end of each

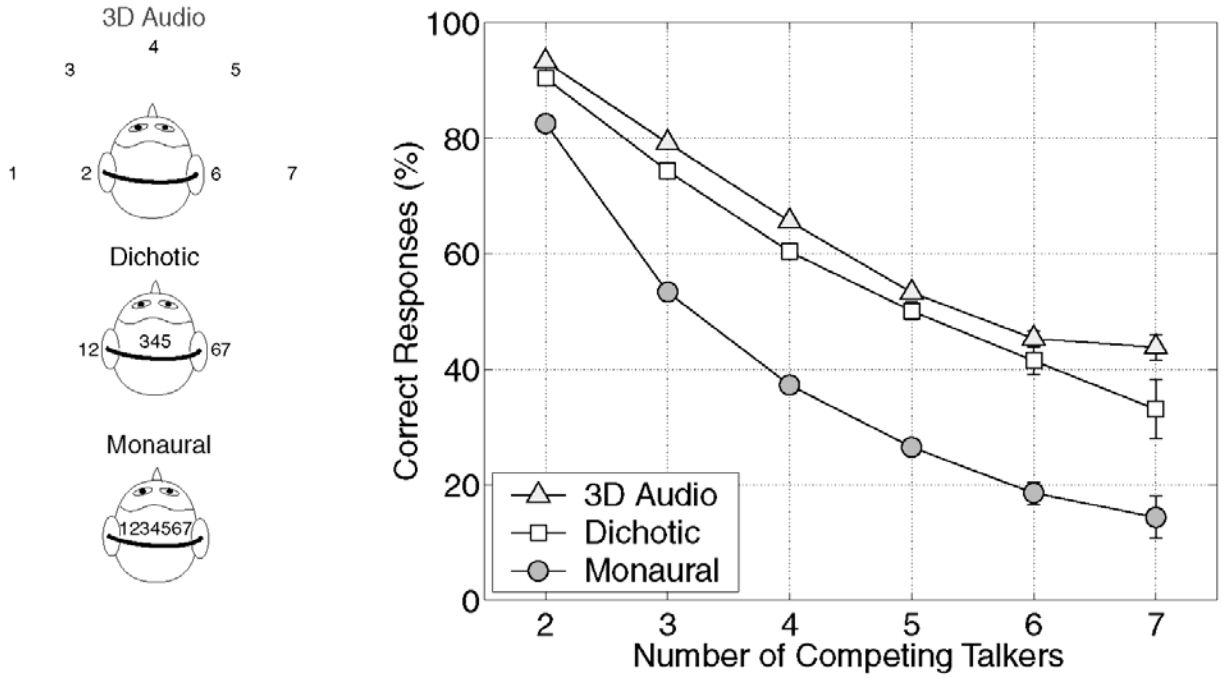

Figure 7: Comparison of overall performance with spatialized, dichotic, and monaural audio displays in a seven-talker CRM listening task with all male talkers. Each talker was 50% likely to be active in each trial, and there was a 25% chance that the target phrase would move to a different talker location at the end of each stimulus presentation. The results have been plotted as a function of the number of active talkers in the trial. The error bars represent the 95% confidence intervals for each data point.

stimulus presentation. The results demonstrate that the spatialized audio display improved performance by about 30 percentage points for all the listening situations with more than two competing talkers. They also show that the 3D Audio display consistently improved performance by about 5 percentage points relative to the dichotic speech display, even in situations where only a small number of the speech channels were active. This seems to confirm that the additional expense of upgrading a dichotic speech display into a spatialized speech display is justified in situations where accurate communications are necessary to ensure successful accomplishment of the mission.

4.0 CONCLUSIONS

In the design of military intercom systems intended to accommodate more than one simultaneous channel of speech communications, virtual synthesis techniques that spatially separate the apparent locations of the competing talkers are an effective and affordable way to improve communications efficiency and enhance warfighter performance. The data from Figure 5 show that spatially separating same-sex competing talkers by 45° produced a 25-35 percentage point increase in overall performance in the CRM task. In terms of the other

factors examined in this paper, this is roughly equivalent to: 1) reducing the number of competing talkers in the stimulus by 1 to 1.5 talkers (Figure 2); 2) replacing the same-sex interfering talkers with different-sex interfering talkers (Figure 3); or 3) increasing the target-to-masker ratio by 3-9 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, while 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 target-to-masker ratio 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 non-speech 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 (Nelson et al., 1999; Drullman & Bronkhorst, 2000). Similarly, real-time headtracking 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 spatialization cues are possible in a system with only one analog output channel, and unfortunately this can be a severe impediment in efforts to attempt to upgrade legacy intercom systems to accommodate spatial audio. However, on the basis of the results presented here, it is clear that spatial audio should be given serious consideration both in the design of new military intercom systems and in the upgrading of existing systems to accommodate evolving C4ISR requirements.

5.0 REFERENCES

Abouchacra, K., Tran, T., Besing, J., & Koehnke, J. (1997). Performance on a selective attention task as a function of stimulus presentation mode. Proceedings of the Midwinter Meeting of the Association for Research in Otolaryngology, St. Petersburg Beach, Florida.

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., Nelson, W., Ericson, M., & Simpson, B. (2000). A speech corpus for multitalker communications research. Journal of the Acoustical Society of America, 107, 1065--1066.

Bolia, R. S., Nelson, W. T., & Morley, R. M. (2001). Asymmetric performance in the cocktail party effect: Implications for the design of spatial audio displays. *Human Factors*, *43*, 208–216.

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 Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29-August 1, 2001, pp. 169--174.

Brungart, D. & Simpson, B. (2002). The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *Journal of the Acoustical Society of America*, *112*, 664–676.

Brungart, D.S., Simpson, B.D., Kordik, A.J. and McKinley, R.L. (2003). The impact of headtracking on intelligibility in a multitalker display. Proceedings of HFES 2003, Denver, Colorado, October, 2003.

Brungart, D.S., and Simpson, B.D. (2003). Optimizing the spatial configuration of a seven-talker speech display. Proceedings of the 2003 International Conference on Auditory Display, Boston, Massachusetts, July 6-9, 2003, pp. 188-191.

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 of masking for speech through interaural time delay. Journal of the Acoustical Society of America, 42, 124--138.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, *25*, 975–979.

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.A., Brungart, D.S. and Simpson, B.D. (2004). Factors that influence intelligibility in multitalker speech displays. International Journal of Aviation Psychology. 14(3), 313-333.

Ericson, M. & 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 N.J.), pp. 701-724.

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

Kryter, K. (1962). Methods for calculation and use of the articulation index. Journal of the Acoustical Society of America, 34, 1689--1697.

MIL-STD-1472E (1997). Department of Defense Design Criteria Standard. Department of Defense.

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.

Moray, N., Bates, A. and Barnett, T. (1964). Experiments on the four-eared man. *Journal of the Acoustical Society of America*, 38, 196-201.

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. Proceedings of the 43rd Meeting of the Human Factors and Ergonomics Society, 1202--1205.

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.

Triesman, A. (1964). The effect of irrelevant material on the efficiency of dichotic listening. *American Journal of Psychology*, *77*, 533–546.

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.

Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. In G. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance, Second Edition*. Portland: Allyn and Bacon.

Yost, W. (1997). ‘‘The cocktail party problem: Forty years later,’’ in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson. Erlbaum, Hillsdale, NJ, pp. 329–348.