AFRL-RH-WP-TR-2009-0032

Sound Localization in Multisource Environments

Brian D. Simpson
Douglas S. Brungart
Nandini Iyer
Warfighter Interface Division
Battlespace Acoustics Branch

March 2009

Final Report for October 2004 to September 2008

Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Acoustics Branch
Wright-Patterson AFB OH 45433

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2009-0032 HAS BEEN REVIEWD AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR

//signed//
Brian Simpson
Program Manager
Battlespace Acoustics Branch

//signed//
Daniel G. Goddard
Chief, Warfighter Interface Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-03-2009	Final	October 2004 – September 2008

4. TITLE AND SUBTITLE
Sound Localization in Multisource Environments

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Brian D. Simpson
Douglas S. Brungart
Nandini Iyer

5d. PROJECT NUMBER
2313

5e. TASK NUMBER
HC

5f. WORK UNIT NUMBER
2313HC51

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Battlespace Acoustics Branch
Wright Patterson AFB OH 45433-7901

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHCB

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2009-0032

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
88 ABW PA Cleared 04/07/09, 88ABW-09-1363.

14. ABSTRACT
Although most sound localization research has examined the ability of listeners to determine the location of single sounds presented in quiet (typically anechoic) environments, most real-work listening situations are more complex, with multiple simultaneous sounds. Similarly, many applications of spatialized auditory (3D audio) displays are likely to require the presentation of complex auditory virtual environments, which must be reliably perceived and interpreted. Moreover, these displays must function properly even in real-world environments that are often much harsher than the laboratory environments in which they were first developed (e.g., the Warfighter in a battlespace).

15. SUBJECT TERMS
Auditory attention, auditory displays, auditory segregation, concurrent sound sources, FY05 AFOSR, sound localization

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Brian D. Simpson
U	U	U	SAR	38	19b. TELEPONE NUMBER (*Include area code*)

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39 18

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

PREFACE

Experience in real-world listening situations suggests that human observers have the ability to monitor spatial information about multiple simultaneous, or nearly simultaneous, sound sources in a complex auditory environment. Such situations may involve 'hearing out' a single talker among multiple, spatially-separated talkers (the 'cocktail-party' effect; [Cherry, 1953]) or one in which listeners must determine the specific location of a sound source in a complex, multisource scene. Although the cocktail-party effect has received considerable attention in the last 50 years, very little research has addressed multisource sound localization. While most of our understanding of spatial hearing comes from relatively simple stimulus situations with single sound sources in anechoic environments, environments in the real world, as well as those where auditory displays are incorporated, are much more complex, and the study of multisource localization is critical to understanding behavior in these environments. Moreover, data from multisource sound localization experiments can provide critical tests for models of spatial hearing. The goal of this program of research was to systematically examine sound localization in multisource environments. Because there is relatively little directly-related prior research in this area, a broad approach was taken, including examining the detection, identification, and localization of a single known source, the "target," when other sources were present; the localization, when cued (either before or after the fact), of an arbitrary source in a multisource environment, and the ability to selectively attend to multiple sources when additional irrelevant (but presumably interfering) sources are present. As this effort was an attempt to bridge the gap between performance in laboratory environments and experiences in real worlds, much of the work emphasized more ecologically-valid environmental sounds (e.g., speech, animal sounds, vehicle noise, etc.) but simple sounds (clicks and noise) were also employed.

THIS PAGE INTENTIONALLY LEFT BLANK

1.0 Introduction

Although most sound localization research has examined the ability of listeners to determine the location of single sounds presented in quiet (typically anechoic) environments, most real-world listening situations are more complex, with multiple simultaneous sounds. Similarly, many applications of spatialized auditory (3D audio) displays are likely to require the presentation of complex auditory virtual environments, which must be reliably perceived and interpreted. Moreover, these displays must function properly even in real-world environments that are often much harsher than the laboratory environments in which they were first developed (e.g., the warfighter in a battlespace).

McKinley and Ericson (1997) considered a number of applications of 3D audio for fighter cockpits, including spatialized communication channels, navigation aids, target indicators, and threat warnings. These spatialized signals have to be presented such that they are reliably localized and understood, along with other cockpit warnings and direct auditory indices of the aircraft status, in the face of considerable engine noise (e.g., approximately 110 dBA in an F-16). Auditory displays for the ground soldier face similar problems. Multiple targets will need to be simultaneously and efficiently rendered in the presence of substantial background noise. In an urban combat situation, spatialized auditory signals could indicate the positions of squad members, including a fallen comrade; the dimensions and paths of safe ingress and egress for a smoke-filled room; and target and threat locations. Again, these virtual signals need to afford one the ability to hear subtle real-world sounds (e.g., the footsteps of an enemy) and must function despite vehicle noise, gunfire, explosions, etc., arising from the surrounding battle.

The few laboratory data on sound localization in the presence of multiple sources suggest that sound localization performance is likely to degrade substantially when more than one sound is presented simultaneously (Good and Gilkey, 1996; Langendijk et al., 2001; Lorenzi et al., 1999; Simpson, 2002). These findings are in sharp contrast to everyday listening experiences in which it appears that listeners have considerable knowledge of the spatial relations of multiple concurrent sounds. Although this ability clearly involves top-down as well as bottom-up processing, and serial as well as parallel processing, this experience suggests that considerable spatial and semantic information can be gleaned from a sufficiently realistic spatial display. Similarly, observers listening to binaural recordings routinely report clear spatial percepts for multiple simultaneous sources. Moreover, binaural recordings, even though they lack critical elements such as individualized head-related transfer functions (HRTFs) and tracked head movements, are typically described as sounding better than signal-processing-based auditory displays, with better externalization, less localization blur, and more realism. Compared to the stimuli typically employed in the laboratory, binaural recordings are more like real-world listening in that they include realistic environmental sounds, echoes, and reverberation. The result is an ordered, but complex, spectral/temporal pattern of stimulus information. On the surface it appears that a listener in the real world is able to analyze this pattern into individual sources, and associate at least some spatial information and some semantic information with most of the sources.

The goal of this effort was to explore the characteristics of multisource spatial hearing and to determine what spatial information a listener can glean about a multisource environment.

In some listening situations, a listener's primary task was to hear out a particular sound and ignore the rest; other situations require listeners to attend to more than one sound simultaneously in order to perform a task. These two situations were examined in the studies described in this report. In these listening situations, it was reasonable to ask what information the listener had about the primary source (a question related the ability of a listener to selectively attend to a specific sound) and also what information the listener had about those sources not the focus of selective attention (a question related to the ability of a listener to divide attention across multiple sounds) (Cherry, 1953; Brungart and Simpson, 2002). These are extreme cases in which listeners must attempt to either ignore most of the sources or attempt to listen to all of the sources. It is reasonable to assume that the differences in processing in these situations lie primarily at the cognitive level and that the underlying sensory processes are highly correlated across the two situations. And so, it was anticipated that information gained from one situation would provide insight about the processing used in the other situation. Of particular interest here was a phenomenon related to the monitoring of multiple streams of auditory information. Specifically, because the auditory system receives input from all locations in a listener's immediate environment continuously, it is believed that a listener at least loosely monitors the location of many sound sources often without consciously attempting to localize them. This monitoring serves a critical survival function (situation awareness) and helps to maintain a sense of being in the world (Gilkey and Weisenberger, 1995; Ramsdell, 1978). Obviously, this type of monitoring is of great interest, and has great potential utility in auditory displays.

Although it appears that, in most listening situations, observers have considerable information about the spatial aspects of the auditory environment around them, they may have difficulty reporting this information in a typical psychoacoustic experiment. The failure of laboratory studies to reveal the level of performance that might have been anticipated based on everyday listening is likely to have resulted in part from the measurement techniques that have been used. It is critical to identify procedures that readily allow the subject to report the content of their spatial perceptions. A number of issues need to be considered. If the subjects are unable to identify or name the individual sounds, it may be difficult for the subjects and/or the experimenters to associate a particular localization judgment with a particular sound source. Therefore, this effort included some examination of various tasks needed in order to choose efficient and appropriate data collection methods. We conducted a series of experiments to identify some response tasks that could both be reliably performed by the subject and had psychophysical utility and/or operational relevance.

2.0 Experiment 1: Detection and Localization of Speech in the Presence of Competing Speech Signals

2.1 Introduction

Auditory displays have been employed in a variety of applications, from simple alarms and warnings in automobiles to advanced virtual audio display technologies in aircraft cockpits. A common issue in the design of these displays is the tradeoff between the desire to present the listener with as much information as possible and the concern that the listener will be unable to process and interpret the auditory information if too many sounds are presented at the same time. This can be a particularly important issue in speech-based auditory displays that present information via prerecorded voice samples rather than more abstract sounds. This paper presents the results of an experiment that evaluated listeners' ability to detect and localize speech-based audio tokens in a display where multiple competing tokens are presented at the same time.

While many types of auditory displays could potentially be used to present multiple simultaneous warning sounds, we decided to focus initially on speech displays. Speech displays have the advantage that they are intuitive and thus can be understood with little or no training on the part of the operator. In addition, they lack the ambiguity that so often typifies many nonspeech auditory symbologies and they can be used to convey almost any kind of information. However, there are a number of potential disadvantages to using speech as the basis for an auditory display. First, speech intelligibility can degrade rapidly in noisy environments (Miller, 1947), which can result in an operator misinterpreting or completely missing a critical signal. Whereas such difficulties may be overcome in nonspeech displays by careful manipulation of the stimulus parameters to accommodate such environments without distorting the meaning of the stimulus, such is not necessarily true in the case of speech. Another disadvantage of speech is that most of the energy in speech signals is concentrated in the lower frequency region (i.e., below 6 kHz), which means that speech signals may lack the high-frequency information needed to support accurate sound localization, particularly in regards to elevation determination and front/back discrimination (Gilkey and Anderson, 1995). This issue is important because the ability to convey spatial information independent of the semantic content of a speech stimulus is desirable for future spatial auditory displays in which the location of the speech signal itself may convey critical information.

A possible problem with the use of speech displays is that the listener may be unable to extract information from the most relevant auditory warning when more than one warning sound is presented at the same time. Such warning sound "collisions" can result in display stimuli that are distorted or obscured, and this can lead to reduced detectability of critical signals, lowered recognition rates, and a general degradation of stimulus localizability. Despite the importance of these issues, the guidelines employed for implementing speech-based auditory displays have traditionally relied on laboratory research, most of which has

employed relatively simple stimulus situations in which a single source or small number of sources are presented simultaneously. Little is known about the detectability and localizability of speech in the presence of a large number of competing speech phrases. The goal of this study was to examine both the detection and localization of a speech signal as a function of the number of sources present and the relative locations of these sources.

2.2 Methods

2.2.1 Participants

A total of 7 paid volunteer listeners (3 males and 4 females, 20-25 years of age) participated in the experiment. All had normal hearing (i.e., bilateral thresholds < 15 dB HL from 125 Hz to 8000 Hz) and all had participated in previous experiments involving both detection and localization.

2.2.2 Apparatus

The Auditory Localization Facility in the Air Force Research Laboratory at Wright-Patterson Air Force Base was used for the collection of behavioral data. This facility consists of an anechoic chamber, the walls, floor, and ceiling of which are covered with 1.1-m thick fiberglass wedges to reduce echoes. A 4.3-m geodesic sphere (see Figure 1), which has 277 Bose 11-cm Helical-Voice-Coil, full-range loudspeakers mounted on its surface, is housed in the chamber. The loudspeakers that were utilized in this study (239 in total) surrounded the listener (360 in azimuth and from -45 to +90 in elevation) and were directed toward the listener's head, which was positioned at the center of the sphere. (Those loudspeakers below -45 U/D were not utilized in this experiment because the direct path to the listener from these loudspeakers was, in some cases, obstructed.) This large set of locations reduced the potential for a listener to make categorical, rather than absolute, localization responses, as may be the case when more restricted sets of sound source locations are tested. Mounted directly in front of each loudspeaker on the sphere is a square cluster of four LEDs.

2.2.3 Stimuli

The auditory stimuli employed in this experiment were 50 phonetically balanced (PB) monosyllabic words drawn from a single list of the PB50 word list corpus. This list was spoken by each of 12 talkers (6 male and 6 female) for a total of 600 unique speech tokens. The speech tokens were broadband (.2kHz - 16kHz), and were level normalized. They were also processed with the Pitch Synchronous Overlap and Add (PSOLA) algorithm in PRAAT to change their durations to exactly 500 ms.

On each trial, a target was defined by a specific speech token (i.e., a specific word spoken by a specific talker). On target-present trials, the target speech token was accompanied by the presentation of between 0 and 5 competing speech tokens. Relative to the target speech, each competing speech token was spoken by either the same talker, a different talker but of the same sex, or a different-sex talker. On target-absent trials, between 1 and 6 non-target

Figure 1: The Auditory Localization Facility at Wright-Patterson Air Force Base. See text for details.

speech tokens were presented. The individual talker characteristics were similar to those in target-present trials (i.e., all same talker, all same sex, or 1 talker that was a different sex than the other talkers), and the speech tokens were selected such that one of the tokens came from the target talker.

The individual speech tokens were convolved with the inverse transfer function from the appropriate loudspeakers in order to remove the effects of the loudspeaker frequency responses, and were then sent from an experimental control computer to a Mark of the Unicorn (MOTU 24 I/O) digital-to-analog converter. Each signal was then sent to a separate channel from a bank of power amplifiers (Crown Model CL1). These amplified signals were then directed to a custom-built signal-switching system (Winntech) before each individual signal was routed to the appropriate loudspeaker. On half of the trials, the speech tokens were spatially separated from one another, with the constraint that the angular separation between all active loudspeakers was at least 45 degrees (the 'spatially separated' condition), and on half of the trials all speech tokens were presented from the same loudspeaker (the 'co-located' condition).

2.2.4 Procedure

During the experiment, each listener stood on a platform in the middle of the Auditory Localization Facility. The listeners' task was to determine whether or not a particular speech token was present (the detection phase) and then, if present, to determine the location of that speech token (the localization phase). At the start of each block of trials, the listener was required to turn to face a reference loudspeaker located directly in front of her/him on the horizontal plane and boresight a hand-held tracking device (the 'wand'; Intersense IS900),

which was subsequently used to record both detection and localization responses. An LED cluster, co-located with this reference loudspeaker, was then activated briefly to indicate the start of a trial. This was followed by a cuing interval, during which the target speech token was presented (Note: in order to avoid biasing the listeners' localization responses with a directional cue, the cued target speech token was presented from the 4 horizontal-plane polar loudspeaker locations simultaneously, resulting in a diffuse image). A subsequent 500-ms silent interval was followed by the observation interval, during which the stimulus (between 1 and 6 simultaneous speech tokens) was presented.

The listener first judged whether the target was present or absent. If the target was judged to be present, the listener was required to indicate the perceived location of the target by pointing the wand at the appropriate loudspeaker and pressing a button; the orientation of the wand was indicated by activating the LED cluster at the loudspeaker to which the listener was pointing (i.e., the wand served as an LED 'cursor'). Note that, on these trials, this single localization response also served as a positive detection response. If the target speech token was judged to be absent, the listener depressed a button on the wand to indicate a 'target-absent' response. If, however, the target was present but was judged to be absent (i.e., a 'miss'), the listener was nevertheless required to make a localization response. No constraints were imposed on head movements throughout the trial, but the listener was required to re-orient to the reference loudspeaker before the start of each subsequent trial. Trial-by-trial feedback was provided regarding the correctness of the detection response and the true location of the target speech token.

In each block of 48 trials, 2 trials were run in each combination of number-of-competing tokens (0-5), spatial configuration (spatially separated and co-located) and target state (present or absent). The *a priori* probability of a target-present trial was 0.5. Only one talker characteristics condition (same-talker, same-sex, different-sex) was tested in each block, and 16 blocks were run in each of these conditions, for a total of 48 blocks per listener.

2.3 Results and Discussion

As would be expected, all of the curves in Figure 2 show that the listeners were able to correctly detect the presence of the target speech token 100 percent of the time when it was the only token presented. It can also be seen that overall detection performance decreased as the number of simultaneously presented competing speech tokens increased. However, the rate at which detection performance decreased was remarkably slow. Even in the worst case tested, where the target speech token was presented in the context of five simultaneous competing speech tokens spoken by the same talker in the same location (gray triangles in righthand panel), listeners were able to correctly detect the presence of the the target more than 70 percent of the time. This suggests that the detection of a known monosyllabic target word in the presence of simultaneous masking words is a remarkably robust process that may be possible even in very adverse listening environments containing multiple similar sounds.

Comparing the different curves within each panel of the figure, it is apparent that similarity between the target voice and competing voices does have an impact on the ability to detect the target. When the stimulus contained four or five competing speech tokens,

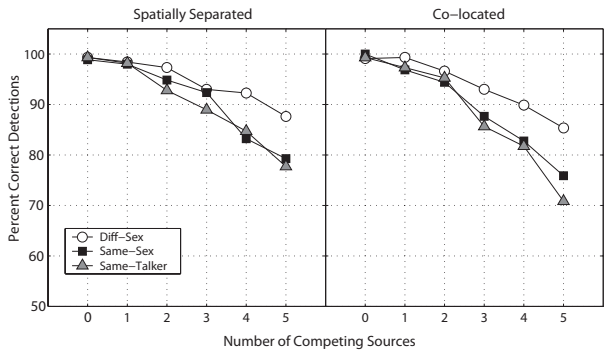

Figure 2: The percentage of correct detections plotted as a function of the number of simultaneous competing speech tokens. The lefthand panel depicts the data from the trials in which the speech tokens were spatially separated from one another (the spatially separated condition), and the righthand panel depicts the data from trials in which all speech tokens emanated from the same loudspeaker (the co-located condition). The parameter in each panel is the specific talker characteristics condition tested (different sex target condition, same sex target condition, same talker target condition).

detection performance was consistently 8-10 percentage points better when the target voice was different in sex than the competing tokens (open circles) than when it was the same sex as the competing tokens. On the surface, one might attribute this difference to the fact that the listener in the different-sex condition only needs to listen for the presence of a talker of a particular gender (e.g., a female voice in the presence of male voices) rather than for the actual key word spoken by that talker. However, the stimuli in this experiment were balanced so that the target-absent trials in the different-sex conditions contained the same mix of genders as the target present trials (for example, one female talker and five male talkers in the six-talker condition) and always contained a speech token from the cued talker. Thus the greater detection performance obtained for the difference-sex condition, shown in Figure 2, cannot be attributed to a detection strategy based solely on the recognition of a female target in the presence of male maskers. The most likely explanation is that the listeners in the different-sex condition were able to immediately focus their attention on the word spoken by the odd-sex talker in the stimulus, and that this made it substantially easier for them to determine if the word spoken by that odd-sex talker matched the cued target token.

Comparing the left and right panels of Figure 2, we see one of the most surprising results of the experiment: the listeners performed nearly as well in the co-located condition as they did in the spatially separated condition. More specifically, performance in the co-located condition was sufficiently good such that very little additional release from masking was seen when the tokens were spatially separated. These results appear to be inconsistent with previous results in the literature demonstrating that spatial separation does, in fact, yield improved detection performance (Good et al., 1997) and speech *intelligibility* (Drullman and Bronkhorst, 2000). However, the results are in fact consistent with the notion that the spatial release from masking is very small when performance in the baseline condition (in this

case, the co-located condition) is sufficiently good (Hirsh, 1948). A closer look at the data, however, indicates that detection performance in the co-located condition degrades more rapidly than performance in the spatially separated condition as the number of competing speech tokens increases. That is, the spatial release from masking is increasing as the number of competing sounds increases. This trend suggests that much larger releases from masking might be found if the number of competing sounds extended beyond 6.

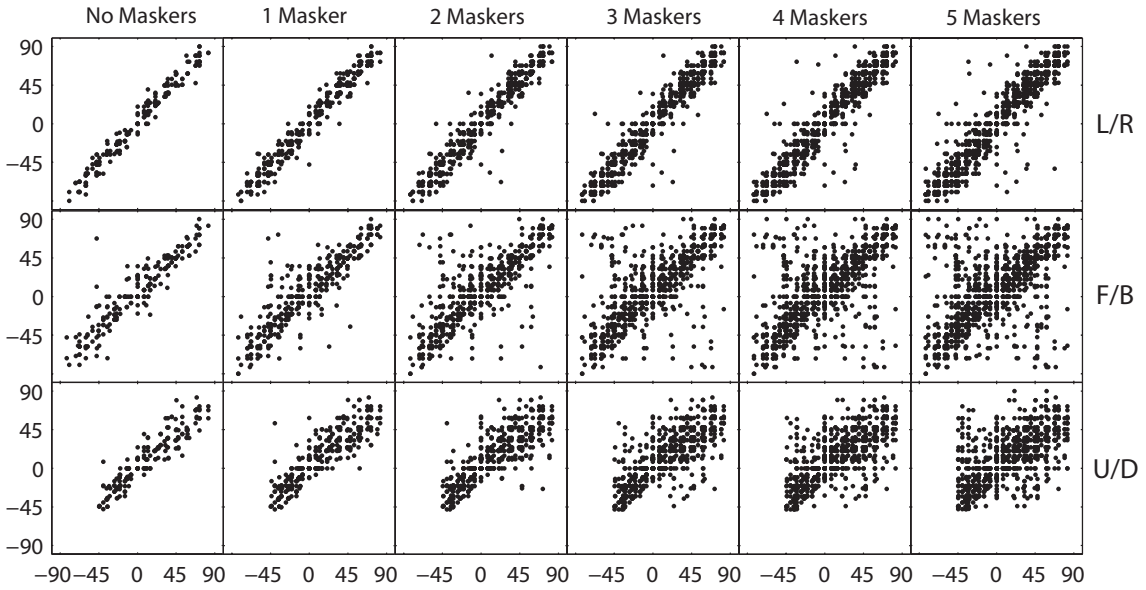

Figure 3: Localization responses plotted as a function of the actual source locations for all listeners in the left/right dimension (top row), front/back dimension (middle row) and up/down dimension (bottom row). The number of competing sources increases as you move from the left-most panel to the right-most panel. Perfect performance would result in all responses falling along the positive-slope diagonal.

The results from the localization task are shown in Figure 3 for all listeners in all cases where the speech tokens were spatially separated and the target was correctly detected. Each row depicts the data for a single spatial dimension (left/right, L/R; front/back, F/B; up/down, U/D), as the number of competing talkers varies from zero (the left-most panel) to 5 (the right-most panel). As can be seen, localization in the L/R dimension was found to be quite accurate, as can be seen by the proximity of the data points to the positive-slope diagonal, particularly when the number of competing sources was small. Localization in the U/D dimension was worse than the L/R dimension, as indicated by a greater spread of data points around the positive-slope diagonal. Localization in the F/B dimension was worse than both the L/R and U/D dimensions. These results are consistent with previous results in the literature (e.g., (Wightman and Kistler, 1997)). One can also see that, as the number of competing sources increases, localization accuracy degrades systematically in all dimensions, but much more rapidly and to a much greater extent in the F/B and U/D dimensions. These results are summarized in Figure 4, where the mean RMS errors in each spatial dimension are plotted as a function of the number of competing speech tokens. In

all dimensions, the RMS errors increase with the number of competing sounds. However, the errors in the L/R dimension remain relatively low, not exceeding 18 degrees until more than four competing sounds are present in the stimulus. The RMS errors are slightly larger in the U/D dimension, and are larger still in the F/B dimension. In fact, the RMS errors in the F/B dimension are greater at every point along the curves than those in the L/R and U/D dimensions for the corresponding conditions. It is interesting to note that the similarity between the target voice and the voices of the competing speech tokens makes no difference in the F/B and U/D dimensions, but that localization in the L/R dimension does, in fact, seem to be better when the target is a different-sex than when it is more similar to the competing voices.

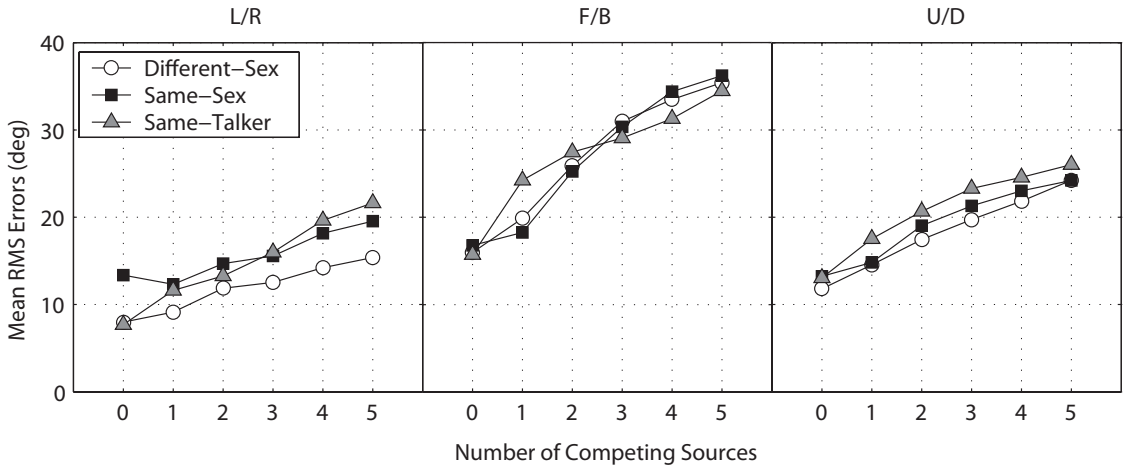

Figure 4: RMS errors, averaged across all listeners, are plotted as a function of the number of competing sources in the L/R, F/B, and U/D dimensions.

Figure 5 combines the L/R, F/B, and U/D localization errors shown in Figure 3 into a single overall measure of angular (great circle) error. As in Figure 2, the two panels show performance in the two spatial conditions of the experiment, and the individual curves within each panel show the different target-masker similarity conditions in the experiment. In the easiest localization conditions, where the target token and/or competing tokens were all presented from the same spatial location (i.e. the no-masker condition in the left panel of the figure and all co-located conditions in the right panel of the figure), the overall angular errors averaged approximately 15 degrees. Note that this is roughly the same angular error reported by (Wightman and Kistler, 1989) for broadband sounds. In part, the relatively high level of performance obtained for the speech stimuli in this experiment can be explained by the use of some exploratory head movements. The 500 ms stimulus duration in this experiment was not long by any means, but it probably afforded the listeners some opportunity to initiate a head movement and thus helped to reduce front-back confusions. The front-back confusions and elevation errors were probably also reduced by the use of broadband speech recordings. Recent studies have shown that sufficient high-frequency information is preserved in broadband speech to support relatively accurate localization (Best et al., 2005), despite

the fact that most of the energy (and virtually all of the intelligibility information) in speech is concentrated at frequencies below 6 kHz.

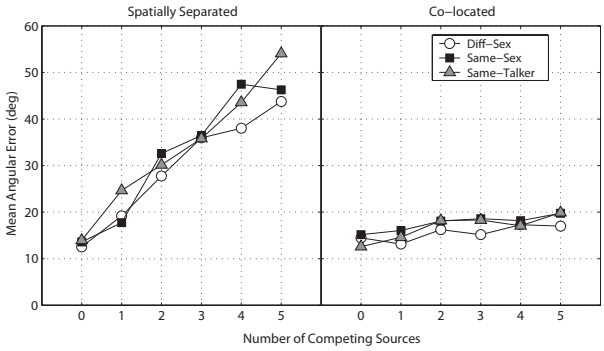

Figure 5: Overall angular errors in the experiment plotted as a function of the number of interfering sounds. These data are plotted in the same format used in Figure 2.

As the number of maskers in the spatially separated condition increased (left panel), the average localization error increased in a nearly linear fashion, with approximately a 5-7 degrees increase in angular error for each additional masker added to the stimulus. However, it is worth noting that performance remained well above chance performance (approximately 90degrees error) even in the worst case tested with five simultaneous maskers. In that case, the overall average error was around 45 degrees, which, although not very accurate, does indicate that the listeners were able to recover some spatial information about the target.

2.4 Conclusions

Listeners' ability to detect and localize a target speech token was measured as a function of the number of competing speech tokens and the spatial separation among these tokens. The results show that although performance decreased as the number of competing sources increased, both detection and localization where surprisingly accurate even with 5 competing sources. Additional research is needed to examine how performance degrades when even greater numbers of sources are used, to determine the role of head movements, and to reconcile apparent inconsistencies with previous "cocktail-party" effect experiments. Of particular interest is the functional relation between detection and localization mechanisms. In this study where the target token is known (via the cuing interval), but the target location is not, spatial separation has little impact on detection performance, apparently supporting a "what-then-where" strategy. This hypothesis could be systematically examined in a study that varied the uncertainty of the target token and the target location.

3.0 Experiment 2: Comparison of Pre-Cueing and Post-Cueing in Multisource Localization

3.1 Introduction

It is assumed that in most listening situations observers have considerable information about the spatial aspects of the auditory environment around them, but that they may have difficulty reporting this information in a typical psychoacoustic experiment. The failure of laboratory studies to reveal the level of performance that might have been anticipated based on everyday listening is likely to have resulted in part from the measurement techniques that have been used. It is critical to identify procedures that readily allow the subject to report the content of their spatial perceptions. A number of issues need to be considered. If the subjects are unable to identify or name the individual sounds, it may be difficult for the subjects and/or the experimenters to associate a particular localization judgment with a particular sound source. ....some examination of various tasks is needed in order to choose efficient and appropriate data collection methods. In this experiment, psychophysical procedures are employed that involve identifying a target stimulus before the overall multisource stimulus is presented (the 'pre-cue' condition), or identifying the target stimulus after the overall multisource stimulus has been presented (the 'post-cue' condition).

3.2 Methods

3.2.1 Listeners

Eight listeners, ranging in age from 19-29 years, served as listeners. All had audiometric thresholds within the normal range (< 25 dB HL at octave frequencies between 250 - 8000 Hz) and were well-practiced in similar listening tasks. Subjects were paid for their participation in the experiments.

3.2.2 Apparatus

As in the previous experiment, Experiment 2 was conducted in the Auditory Localization Facility (ALF) in the Air Force Research Laboratory at Wright-Patterson Air Force Base (see Figure 1).

3.2.3 Stimuli

Environmental sounds (e.g., telephone, tire screech, applause, etc.) and speech tokens (50 words from the phonetically-balanced (PB) word lists spoken by each of 12 talkers (6 m, 6 f) for a total of 600 unique speech tokens) were employed as stimuli. All tokens within a trial were spoken by the same talker. The bandwidth of the stimuli was .2 kHz - 16 kHz, and all were normalized for level. They were also convolved with the inverse of the loudspeaker

transfer function in order to equalize for individual loudspeaker responses. The sounds were normalized to 500 ms using the PSOLA software algorithm, and the stimuli were presented normally or time-reversed. The minimum separation between concurrent sources was 45 degrees.

3.2.4 Procedure

The target stimulus to be localized was identified for the listener by a pre-stimulus or post-stimulus cue. The cue was presented from multiple locations simultaneously (0, 90 -90, 180 degrees azimuth on the horizontal plane + directly overhead) in order to generate a diffuse image and reduce potential biases associated with the cue. The listener's task was to point to the judged target location with a hand-held tracking device that activated LEDs at the selected loudspeaker. The listener was required to re-orient to a loudspeaker directly in front (0 degrees azimuth) after each trial. Trial-by-trial feedback indicating the actual target location was provided.

3.3 Results

The results are shown in Figure 6. Here, it can be seen that, not surprisingly, the average angular error increases as function of number of sources. Moreover, localization errors were found to be lower in pre-cue condition than in post-cue condition. This difference across conditions could be seen with as few as two simultaneous sources. Finally, the errors were found to vary with stimulus type, although this was true primarily in the post-cue condition.

To more closely examine the nature of these localization errors, the data were transformed into a 3-pole coordinate system (Left/Right, Front/Back, Up/Down), as depicted in Figure 3.2.4. These transformed data were subjected to a 3 (spatial dimension) $\times$ 5 (number of sources) $\times$ 2 (cueing) $\times$ 2 (stimulus type) $\times$ 2 (temporal order) repeated measures analysis of variance (ANOVA). In general, localization in the left/right dimension was best; performance was slightly worse in the up/down dimension, and worse still in the front/back dimension, consistent with previous results. Localization in the pre-cue condition was better than that in the post-cue condition in all dimensions. Differences were most pronounced in the left/right dimension, and became larger as the number of sources increased. There was a significant cueing $\times$ number of sources $\times$ dimension interaction [$F(8, 56)=19.76$, $p < .05$]. However, it is important to note that in all conditions and across all dimensions, listeners always performed better than chance in all dimensions. The localization of speech tokens was somewhat worse than that for environmental sounds, but only in the left/right dimension. This effect was found to become more pronounced as the number of sources was increased. There was a significant stimulus type number of sources $\times$ dimension interaction [$F(8, 56)=10.49$, $p < .05$]. The most notable differences between forward and time-reversed stimuli were found in the left/right dimension, but these differences were small, and there was not a significant temporal order $\times$ number of sources $\times$ dimension interaction. Post-hoc

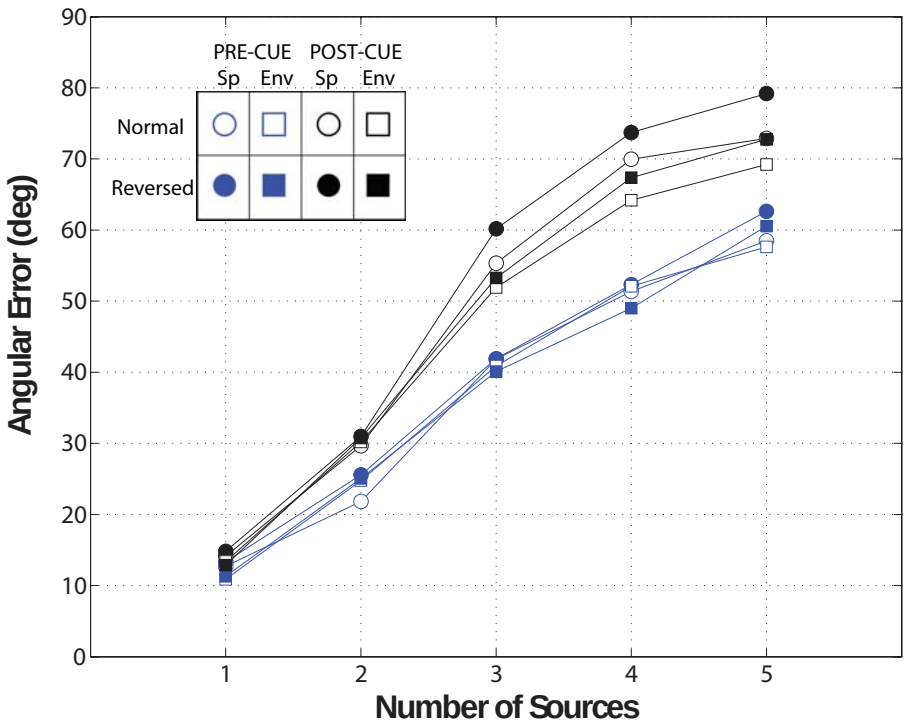

Figure 6: Overall angular localization errors plotted as a function of the number of sources for each stimulus type. The blue symbols represent date from the Pre-Cue condition; black symbols represent data from the Post-Cue condition. Open circles represent the data when normal speech stimuli are employed, and filled circles represent the data when time-reversed speech stimuli are employed. Similarly, open squares represent data for the case in which normal environmental sounds are used, and filled squares represent the case in which the environmental sounds are presented time-reversed.

analysis revealed that the cueing $\times$ temporal order interaction was not significant.

3.4 Discussion

The fact that the listeners did worse in the post-cue condition than in the pre-cue condition suggests that they were not able to localize all of the stimuli simultaneously. Although the intention of the post-cue was to limit the impact of memory by allowing the listeners to make a single response rather than recite the locations of each of the stimuli individually, following Sperling (1960), it cannot be said for certain that memory was not still an issue. In addition, the differences across stimulus type suggest that similarity and discriminability play an important role in multiple-source localization performance. Environmental sounds were more different from one another along spectral/temporal dimensions than were the individual speech tokens within a trial. It seems likely that these differences not only made the

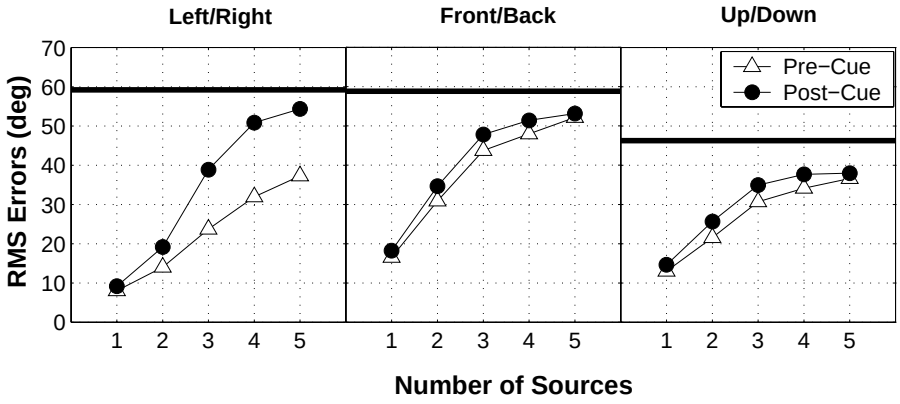

Figure 7: RMS errors plotted as a function of the number of sources in the 2 cueing conditions in the Left/Right (left panel), Front/Back (middle panel) and Up/Down (right panel) dimensions. Data are averaged across stimulus type. Black lines indicate chance performance.

stimuli more discriminable per se, but also allowed listeners to recover localization cues from spectral/temporal regions when the energy in the interfering stimuli was low, and this led to somewhat better localization performance, but only in the left/right dimension. Finally, it appears that, in this study, the differences in identifiability created by comparing forward to time-reversed sounds had little effect on localization performance.

3.5 Conclusions

This was a preliminary study examining sound localization in multiple-source environments. The results indicate that, although performance degrades with the number of competing sources, listeners clearly have some information about the spatial location of sounds, even with 5 simultaneous sources - in no case did the listener reach chance performance. We found that *a priori* information about a source enhances localization performance, as shown by the difference between the pre-cue and post-cue conditions. In addition, performance was found to degrade more rapidly in the post-cue condition than in the pre-cue condition, a difference that was found to exist for as few as two simultaneous sounds. The identifiability of the sound did not have a large impact per se, as indicated by the lack of difference in performance between the forward and time-reversed stimulus conditions. Nevertheless, the greater across-stimulus differences present in the environmental sound set (in contrast to the single talker per trial speech sounds) seems to have improved localization performance at least in the left/right dimension.

4.0 Experiment 3: Cueing by Deletion: Localizing the Missing Source

4.1 Introduction

Most of our understanding of spatial hearing comes from experiments conducted in laboratory settings, where simple sounds (e.g., tones, noise) are presented in quiet, anechoic environments. In general, these studies suggest that sound localization performance can degrade substantially when more than one sound is presented simultaneously (Good and Gilkey, 1996; Langendijk et al., 2001; Lorenzi et al., 1999). However, these laboratory results appear to be in sharp contrast to our experiences in the real world, where the auditory environment typically contains multiple concurrent sounds that are non-uniform and dynamic. The impression of listeners in such environments is typically one in which they could, if required, accurately report the location of each of the individual sounds. In fact, it often appears that a listener need not actively attend to any specific elements in the auditory environment in order to maintain an overall awareness of the multiple elements and their relative locations. The question remains: Why does this difference exist? First, the stimuli employed in most laboratory studies are unlike those that occur in the real world (tones/noises, brief durations). Such stimuli, even if discriminable from one another, are difficult to 'label' and this process would seem to be critical for keeping track of multiple sounds. In addition, current techniques for measuring localization may not capture a listeners true multisource localization ability.

Despite the belief that listeners have considerable information about the spatial attributes of multiple sounds in their auditory environment, measuring this in a typical psychoacoustic experiment is nontrivial. One way to test a listener's ability to localize multiple simultaneous sounds is to turn the sounds off and have the listener report the location of each individual sound from the auditory scene. However, echoic and short-term memory limitations may restrict the ability of a listener to sequentially report localization information retrospectively, and the results from such a paradigm would be difficult to interpret. An alternative method, and one that addresses these memory concerns, is to delete one sound from a multiple-source auditory scene and ask the listener to indicate the location from which the sound was deleted. The assumption is that if the listener can consistently report the location of a sound that has been removed from a scene, the listener knew the locations of all of the sounds in that scene.

This experiment employs a 'cueing by deletion' paradigm to examine a listener's ability to localize multiple sounds simultaneously. Both the complexity of the auditory scene (the number of concurrent sounds) and the length of time that all concurrent sounds in the scene were presented prior to the deletion of the target sound are varied.

4.2 General Methods

4.2.1 Participants

Six paid volunteer listeners (3 males and 3 females, 19-24 years of age), participated in the experiment. All had normal hearing (audiometric thresholds < 15 dB HL from .125 kHz to 8.0 kHz), and all listeners had participated in previous sound localization experiments.

4.2.2 Apparatus

As in the previous experiment, Experiment 2 was conducted in the Auditory Localization Facility (ALF) in the Air Force Research Laboratory at Wright-Patterson Air Force Base (see Figure 1).

4.2.3 Stimuli

The stimuli used in this study were 19 naturalistic sounds (e.g., birds chirping, lawnmower, man coughing, bees buzzing, harp) culled from a commercially available compilation of sound effects (Ghostwriters, 1998). These stimuli were selected to maximize the similarity of the sounds along several dimensions, including bandwidth (and thus, presumably, localizability), identifiability, and the naturalness of the sound when repeated (looped). The sounds were filtered to have a bandwidth of 0.2 kHz - 14 kHz and were normalized to have the same overall RMS level. They had a duration of approximately 2 sec (the exact duration was determined by the natural time course of the individual sound that would allow for looping), and were independently looped during stimulus presentation. Onsets and offsets were temporally windowed with 10-ms cosine-squared ramps. The sounds were convolved with the inverse transfer function of the presentation loudspeaker to minimize any effects that might occur due to differences in the individual loudspeaker responses. The target sound was always presented from one of 16 loudspeaker locations on the horizontal plane, spaced roughly every 30 degrees. The distracter sounds could originate from any of the 28 loudspeaker locations on the horizontal plane. Loudspeakers were selected such that sounds were never co-located, but no other restrictions were made concerning the angular spacing among the sounds.

4.2.4 Procedure

The listener's task was to attend to a multiple-source auditory scene for a predetermined observation interval and identify the location of the sound source that was turned off at the end of that interval. This task was performed with the listener standing on an adjustable platform in the middle of the ALF with her/his head at the height of the loudspeakers on the horizontal plane. Before the start of each trial, the head-slaved cursor was enabled and the listener was required to center her/his head by aligning the cursor with a reference loudspeaker located at 0 degrees azimuth and pressing a button on the handheld device. The LED cluster was then turned off to indicate the start of the trial. Then, this LED cluster was activated once again, this time in a rotating pattern, and remained in this state

throughout the duration of the observation interval. During this interval, 1, 2, 4, 6, or 8 environmental sounds were presented simultaneously and looped continuously for one of four possible durations: 2.5, 4.5, 6.5, or 8.5 seconds. At the end of the observation interval, one sound, the target, was turned off, as was the LED cluster at the reference loudspeaker, but the distracter sounds remained on. This 'distracter-only' interval continued until the listener moved her/his head more than 10 degrees in either direction, at which point all sounds were terminated, indicating the start of the response interval. The LED cursor was then re-activated, and the listener was required to orient her/his head to the loudspeaker judged to be the target location and press the button on the handheld device. Listeners were given trial-by-trial feedback by activating the LED cluster and playing the target sound from the correct response location. After each trial, the listener was required to re-orient the cursor toward the reference loudspeaker before the start of the next trial. Listeners' head movements were constrained by tracking the head position, and the trial was aborted if the head moved more than 10 degrees from the reference orientation during the observation interval.

Within each block of 40 trials, 8 trials were run at each of 5 number-of-source conditions (1, 2, 4, 6, and 8). Only one observation interval duration was run in each block, and two blocks were run at each of the four durations (2.5, 4.5, 6.5, and 8.5 sec), for a total of 320 trials per listener, 16 in each condition. Throughout the experiment, target locations were equally distributed across the 16 designated loudspeakers on the horizontal plane, and distracter locations were randomly selected from all 28 locations on a trial-by-trial basis. The experimental conditions were randomized across listeners. Each listener completed at least one training block to become acquainted with the procedure before formal data collection began.

4.3 Results

4.3.1 Experiment 1

For analysis purposes, the azimuthal localization errors were decomposed into a left/right component and front/back component (Kistler and Wightman, 1992). This system is convenient because the cues that mediate localization in each of these dimensions are different, and thus the resulting errors may be attributed to different underlying mechanisms. The left/right coordinate of a sound source is the angle between the location vector and the median plane (the vertical plane that is perpendicular to the horizontal plane and bisects the interaural axis) and is a measure of stimulus laterality. It is believed that performance in this dimension is based primarily on interaural cues.

Mean left/right localization errors were subjected to a 5 (number of sources) $\times$ 4 (observation interval) analysis of variance (ANOVA), revealing significant main effects of the number of simultaneous sources, $F(4, 20) = 124.302$, $p < .05$, and the duration of the observation interval, $F(3, 15) = 5.484$, $p < .05$, as well as a significant number of sources $\times$ observation interval interaction, $F(12, 60) = 2.139$, $p < .05$. These effects can be seen in Figure 8, where mean localization errors in the left/right dimension are plotted as a function of the number of concurrent sounds presented during the observation interval (i.e., before the deletion of the target sound). The parameter in the graph is the duration of the observation interval.

Single-source localization data were collected as a baseline to ensure that the listeners could accurately localize the environmental sounds employed in this study. Note that although these data were collected for each duration of the observation interval, it was anticipated that there would be no difference across conditions. As is evident in Figure 8, this was indeed the case. That is, at least for the conditions examined in this study, single-source localization errors remained the same regardless of the time provided to listen to each stimulus. Note also that this duration-independent performance was true when the number of sources was increased to two. More important, however, was the fact that listeners' single-source localization judgments were quite accurate - they were, on average, able to localize the individual sources to within 3 degrees of the actual location, suggesting that these individual sounds were sufficiently broadband to support good left/right localization.

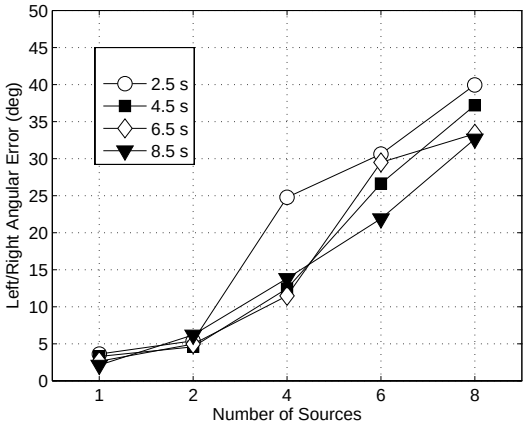

Figure 8: Left/Right localization errors, averaged across all listeners, plotted as a function of the number of sources for each duration of observation interval.

Overall, the data from Figure 8 indicate that left/right localization errors increased as a function of the number of concurrent sources. However, performance degraded differentially depending upon the duration of the observation interval. As stated above, there was little or no effect of observation interval duration when only one or two sources were presented. On the other hand, when the number of sources was four or more, the duration of the observation interval had a substantial impact on localization performance. Specifically, localization errors in the 4-source condition were approximately 11-13 degrees larger (i.e., approximately twice as large) when the observation interval was 2.5 sec than for any other duration. In the 6-source and 8-source conditions, the advantages of a long observation interval were less systematic, but performance was consistently best with the 8.5-sec observation interval, and worst when the listener had only 2.5 sec to hear the auditory scene before the offset of the target. In addition, as can be seen in Figure 9, the proportion of front/back confusions increased systematically with the number of concurrent sources for all durations of the observation interval, but they appeared to do so at a slower rate when the observation interval was the longest. Finally, it is important to note that performance did not vary substantially as a function of the specific sound that was deleted.

4.3.2 Experiment 2

The results from Experiment 1 indicate that the duration of the observation interval could have a substantial impact on a listener's ability to localize the target sound when the number of sources was greater than two. The differences in errors between the 2.5-sec observation interval and the 8.5-sec observation interval were obvious, but the results for the intermediate values were somewhat less clear. Therefore, a second experiment was conducted to more closely examine the impact of observation interval duration on localization. Based on the results from Experiment 1, only a single number-of-sources condition was examined (the 6-source condition), for this was the first condition in which the four durations of the observation interval seemed to differentially impact performance. In order to more fully characterize this impact, two additional durations of the observation interval were included: 1.5 sec and 12.5 sec. Unlike Experiment 1, the duration of the observation interval could vary from trial to trial within a block. In addition, because we were primarily interested in localization performance in the left/right dimension, possible stimulus locations (target or distracter) were restricted to the 16 loudspeakers on the horizontal plane in a listener's frontal hemifield. All other procedures for stimulus presentation and response collection remained unchanged.

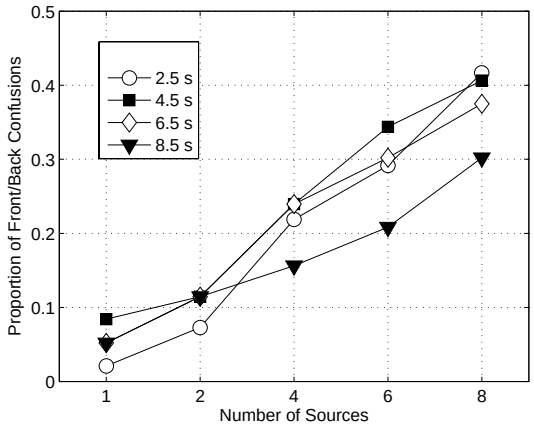

Figure 9: Proportion of front/back confusions, averaged across all listeners, plotted as a function of the number of sources for each duration of observation interval.

The results from Experiment 2 are shown in Figure 10. Here, mean left/right localization errors are plotted as a function of the duration of the observation interval. As can be seen, localization errors decreased systematically as the duration of the observation interval was increased, and a one-way ANOVA revealed a significant main effect of observation interval duration, $F(5, 25) = 12.993$, $p < .05$. When the observation interval was 12.5 sec in duration, mean errors were half as large as those found in the 1.5-sec observation interval condition (15 degrees vs 30 degrees).

Although varying the duration of the observation interval from trial to trial in Experiment 2 introduced uncertainty about when the target would be deleted from the scene, this did not appear to have an impact on performance. Indeed, if we compare the 6.5-sec observation

interval conditions in Experiments 1 and 2, localization errors tended to be somewhat smaller in Experiment 2. This is, perhaps, not surprising if one considers that in the real world, listeners typically have no *a priori* knowledge about when a sound may terminate, yet they are able to determine the location of this event. Moreover, it is possible that keeping the number of sounds constant from trial to trial provided a more stable context against which to judge the location of the target.

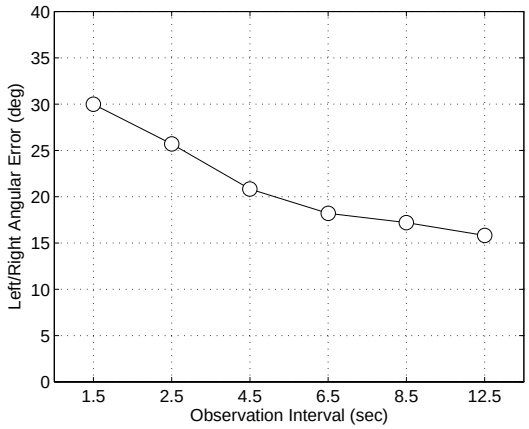

Figure 10: Left/right localization errors, averaged across listeners, plotted as a function of the duration of the observation interval for 6 simultaneous sources.

4.4 Discussion

The results from this study indicate that listeners are surprisingly good at localizing sound in these complex auditory scenes, with localization errors well below chance level of performance in even the most difficult of listening situations. This is particularly impressive given what may be considered a very difficult task - the localization of a sound that is no longer present in the auditory scene. This seems to suggest that listeners were indeed capable of maintaining an awareness of the spatial locations of multiple sources simultaneously.

Although it is the case that the trends found in this experiment are consistent with previously reported results, localization errors in this study were, in general, smaller than those found in previous studies that have required listeners to attend to all of the simultaneous sounds in a multiple-source environment. For example, an earlier study from our laboratory (Simpson et al., 2006) employed environmental sounds to measure localization in multiple-source environments by cueing the target sound either before (pre-cue) or after (post-cue) the observation interval. In the post-cue condition, which presumably required the listener to localize all sounds simultaneously, the left/right localization errors were 15-25 degrees larger than those in the current study under comparable conditions. In part, the larger errors found in (Simpson et al., 2006) can be attributed to the use of much shorter stimulus durations (500 ms). Indeed, even for the pre-cue condition of that experiment, where the target sound was identified prior to the observation interval and the listener was only

required to analytically determine the location of that single sound, left/right localization errors were 5-15 degrees higher than in the conditions in the current study with the same number of sources. This suggests that when complex auditory scenes are presented for short durations, the sounds may simply be more difficult to localize than when they are presented for longer durations, regardless of whether the sounds have to be localized independently or as a group. However, differences in observation interval cannot explain why listeners were able to detect the locations of deleted sources in this study when prior research has shown that listeners, in a similar experimental paradigm, were unable to even *detect* the removal of a sound source from an auditory scene (Eramudugolla et al., 2005) which is presumably a simpler task than localization. This recent study by (Eramudugolla et al., 2005) measured a listener's ability to detect a change between two presentations of an auditory environment and found that listeners were quite poor at detecting these changes unless they were instructed to direct their attention to the item or to the place at which a change might occur. While it is difficult to make direct comparisons between this experiment and the current study, it is the case that listeners in the present study had no information about where to direct attention yet were still able to perform well.

One aspect of the current study that is not shared by the other studies discussed is the fact that a change in the environment is the defining feature of the target stimulus - the stimulus offset - and the listener is exposed to this change. In the earlier studies, a temporal gap was inserted between the stimulus and observation intervals, containing either silence (Simpson et al., 2006) or noise (Eramudugolla et al., 2005). In the current study, listeners may have been able to process changes within a brief integration window to perceive the change, a strategy that would not work for the other studies. Numerous researchers have shown psychoacoustic and electrophysiological evidence demonstrating that changes such as stimulus onsets and offsets may be particularly salient features. However, their salience may depend on the auditory 'background' in which they occur (Gilkey et al., 1990), suggesting that this background provides a context against which to perceive these changes. Moreover, in both (Simpson et al., 2006) and (Eramudugolla et al., 2005) the temporal separation between the stimulus and observation intervals likely allowed for at least some decay of the 'echoic memory trace.' That the duration of exposure to an auditory scene influences a listener's ability to describe a change that has taken place in that scene is wholly consistent with our real world experiences, as well as the data from studies of auditory perception, using noise maskers and tonal signals, which have demonstrated that the duration of masking noise prior to stimulus onset or following stimulus offset (the 'masker fringe') influences stimulus detectability (McFadden, 1966).

Although the results from this study, and those from previous studies, demonstrate that localization performance decreases as the number of concurrent sounds increases, it is not clear to what this decrease in performance can be attributed. It is possible that the increased errors found when the number of concurrent sounds was large results from confusions among, or the summing of, the localization cues from the various sources. That is, a listener may have difficulty segregating these cues associated with the individual sounds and the sum of localization cues from multiple sources would result in ambiguous spatial information. Another possibility is that the reduced signal-to-noise ratio that results from the addition of

competing sounds simply masks the localization cues, rendering them undetectable. Each of these possibilities could lead to a situation in which the listener knew what sound was deleted from the scene but could not discern its location prior to the deletion. A third possibility is that not only are the localization cues masked, but the target sound itself cannot be heard (or is not attended to). In this case, the listener could only make a guess as to the location of the target. Unfortunately, the results from this study cannot distinguish between these explanations. Studies designed to look specifically at the relationship between target recognition and source localization (i.e., between 'what' and 'where') are currently underway in our laboratory.

Finally, it is difficult to determine from these results what strategies the listeners are employing to localize the concurrent sounds. One possibility is that listeners are sequentially 'mapping' the auditory environment, assigning individual sounds to individual locations. Such a process would presumably take time to complete, and the required time might be a function of the complexity of the auditory scene. This would be consistent with the results indicating that more time is required for good localization performance when the number of sources is large. Another possibility is that listeners may tend to listen more 'holistically' to the auditory scene and generate an overall impression, or model, of the spatial layout of the auditory environment - one that does not require attending to the individual sources serially. To the degree that such a model requires time to build up based on the complexity of the auditory scene, this theory is also supported by the data. It is also possible that listeners employ some combination of these strategies, which may vary as a function of the specific listening condition. Based on our current information, it is not possible to distinguish among these possibilities.

4.5 Conclusions

The results from this study clearly indicate that listeners have spatial information about concurrent sounds in a multiple-source auditory scene, and that they can use this information to 'simultaneously' localize these multiple sources. Not surprisingly, this ability appears to vary with the complexity of the auditory scene, as well as the duration of exposure to the scene. Specifically, scenes of greater complexity seem to require more observation time in order to maintain good localization performance. Although in general it seems to be the case that listeners can localize multiple simultaneous sounds in natural scenes, this has nevertheless been a little-researched phenomenon in the auditory literature. Future work will also examine simpler stimuli, including tones and noise, to allow us to systematically identify the specific stimulus properties that lead to effective localization in multiple-source auditory environments.

5.0 Experiment 4: Localization and Identification in Multisource Environments: Does Attention Play a Role?

5.1 Introduction

As previously stated, real world listening experiences suggest that the number of sounds a listener can localize in a multiple-source environment is greater than most results from the laboratory would suggest. The studies described above suggest that listeners can indeed monitor spatial information from multiple sources, although performance was found to degrade systematically as the number of competing sources (the set size) was increased. In addition, performance degraded differentially depending upon the duration of the observation interval. That is, lengthy exposure to the auditory scene appears to provide a context against which a listener may evaluate change. This notion is compatible with studies on the impact of masker fringe on binaural detection (McFadden, 1966; Gilkey et al., 1990).

The previous studies used a method to measure multisource localization called 'cueing-by-deletion,' in which one sound is deleted from a multisource environment, thus designated that sound as the target sound, and the listener's task is then to identify the target sound and/or the location from which the target sound was deleted. The assumption is that if a listener can consistently report the location of the deleted sound, the listener knew the locations of all sounds in the auditory scene. Although listeners performed surprisingly well in the previous studies, several questions remained. First, it was unclear if the listeners actually knew the identity of the target sound when it was deleted, or if they were merely pointing to a location from which they heard a sound turn off. Experiment 4a examined this question by requiring listeners to both identify and localize the target on each trial. In addition, it was not known if the systematic degradation in performance that occurred as the number of sources increased (the so-called 'set-size effect') arose from changes in the signal-to-noise ratio associated with changes in the number of competing sources, or if the effect arose from attentional constraints that limited the number of simultaneous sources to which a listener could attend. Experiment 4b examined this notion by comparing performance when the signal-to-noise ration was held roughly constant by fixing the number of actual sounds presented across trials, but the number of sources to which a listener was required to attend was varied. This subset of sounds within the actual set of sounds was known as the 'Relevant Set.'

5.2 Experiment 4: Methods

5.2.1 Listeners

Six paid listeners (3 male, 3 female),19-24 yrs old, participated in this experiment. All had audiometric thresholds within the normal range ($<$ 25 dB HL at octave frequencies between 250 - 8000 Hz) and were well-practiced in similar listening tasks.

5.2.2 Apparatus

As before, this experiment was conducted in the Auditory Localization Facility (ALF) in the Air Force Research Laboratory at Wright-Patterson Air Force Base (see Figure 1).

5.2.3 Stimuli

The sounds for Experiment 4 were drawn from a subset of 16 environmental sounds from the 19 sounds employed in Experiment 3. As before, the sounds were normalized to be approximately 2 sec in duration. These sounds were selected based on being equally identifiable and equally localizable, as determined by a previous study in our laboratory. All sounds were presented on the listeners horizontal plane.

5.2.4 Procedure

In both tasks of Experiment 4, the listener was required to stand on a platform in the sphere such that his/her head was positioned in the middle of the sphere. The listener's head was to remain fixed during stimulus presentation. A sheet of paper was mounted on a platform in front of the listeners that listed the names of all of the sounds employed in this study.

Experiment 4a Procedure On each trial, 2, 4, 6, 8, 12, or 15 sounds were selected from the list of 16 sounds and presented simultaneously and looped continuously. One sound (the target) was turned off after a period of 2.5 sec or 6.5 sec (the duration of the 'observation interval'), initiating the distracter-only interval. This 'distracter-only interval was terminated when the listener began to orient to the perceived target location. When the listener's head was oriented toward the perceived location of the sound source, the listener pressed a response button on a hand-held wand to record the angle of orientation of their head. In addition, the listener provided a verbal report naming the sound that had been deleted, and this response was recorded by the experimenter in the control room. In this way, localization and identification responses were recorded on the same trials.

Experiment 4a Results

The left panel of Figure 11 shows left/right localization errors plotted as a function of the number of sources, and the right panel shows error rates for target identification plotted as a function of the number of sources. In both panels, data are shown for observation interval durations of 2.5 seconds (open circles) and 6.5 seconds (filled squares). As can be seen, both left/right localization performance and identification performance decrease systematically with the number of sources up to 12 ($p < .001$), after which performance remains the same, suggesting that perhaps there is a floor effect. That is, subjects appear to be performing near chance when the number of sources exceeds eight. Moreover, in both tasks, listeners appear to benefit from a longer observation interval ($p < .05$). That is, localization and identification errors are lower overall when the listeners are provided with a 6.5-sec observation interval as compared to errors when the observation interval is only 2.5 seconds. However, this difference no longer exists when the number of sources exceeds eight. The similarity between the shapes of the curves depicting localization and identification

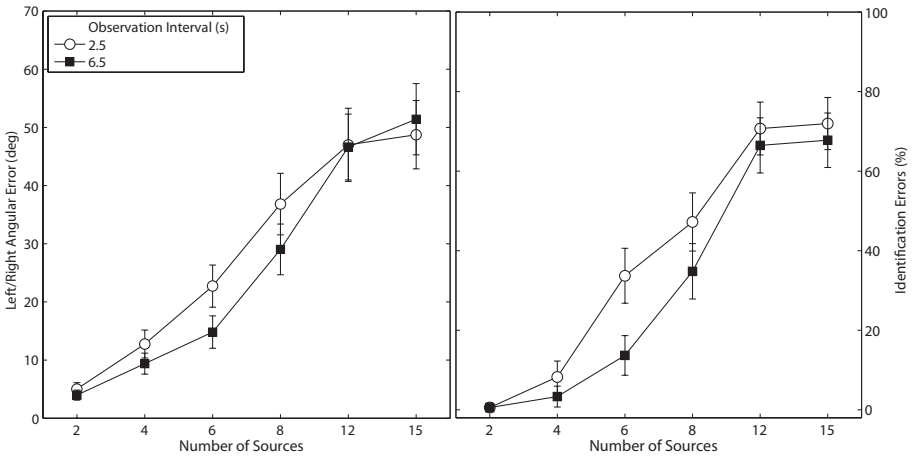

Figure 11: Left/Right localization errors (left panel) and percentage of target identification errors (right panel) plotted as a function of the number of concurrent sources. The parameter in the figure is the duration of the observation interval in seconds. In both panels, error bars represent $\pm$ 1 standard error.

performance suggests that listeners localization performance covaries with their ability to correctly identify the target. That is, it appears listeners were not merely indicating a location from which they heard a sound deleted, but were keeping track of all of sounds.

Although performance was good overall (i.e., better than what would be predicted from previous studies), it was unclear what the limiting factor was in localization performance as the number of sources increased. Specifically, one conceivable explanation is that the decreasing signal-to-noise ratio associated with an increase in the number of competing sounds made the target sound more difficult to localize. This is consistent with the results of (Good and Gilkey, 1996), who found that left/right localization performance began to degrade at low signal-to-noise ratios. Another possible explanation is that listeners are limited in the number of sounds they can 'keep track of' in a multisource localization task. To further examine these possible explanations, another experiment was conducted in which the signal-to-noise ratio was held roughly constant by fixing the number of actual sounds presented across trials, but the number 'relevant' sounds (i.e., a subset of sounds from which one would be deleted) was presented.

Experiment 4b Procedure In Experiment 4b, a total of eight sounds, selected from the 15 sounds used in Experiment 4a, was presented on every trial. Although the specific eight sounds were varied from trial-to-trial, as was the relative location of all of the sounds, it was presumed that on average the overall level of the multisource stimulus, and thus the signal-to-noise ratio, was the same across trials. On each trial, a list of 2,3,4,6, or 8 words associated with the names of the sound sources, was presented (as text) on a computer monitor directly in front of the listener. This list of words designated a set of sounds that were potential targets (that is, only those sounds on the list could potentially be deleted from the multisource scenes). This list was termed the 'Relevant Set.' An additional condition was also run in which no Relevant Set was cued (that is, as in previous studies, the target

could be any of the eight actual sounds presented). In this experiment, the duration of the observation interval was always 6.5 seconds. As before, the observation interval was terminated when the listener began to orient toward the perceived location of the target source, and the head orientation was recorded when the listener pressed a button on the response wand.

Experiment 4b Results Figure 12 depicts left/right localization errors plotted as a function of the number of sources. The open circles represent data when the number of sources refers to the size of the Relevant Set, and the filled squares represent data when the number of sources refers simply to the total number of sources presented on a given trial (replotted from Figure 11). The open diamond represents the case in which no Relevant Set was designated, and thus may be considered to be comparable to the case in which the Relevant Set size is eight (that is, all sources presented are 'relevant'). The dashed line represents hypothetical error values if performance was mediated by signal-to-noise ratio alone (i.e., based on the performance from Experiment 4a when the number of sources was 8 and the observation interval was 6.5 sec). There are a number of things to note from these data. First, localization performance degrades as the number of actual sources increases. This so-called 'set size effect' has been found throughout these studies and is not unexpected. More importantly, this set-size effect is also seen when the actual number of sounds remains constant but the size of the Relevant Set increases. For each 'number-of-sources' condition except 8 (i.e., 1, 2, 3, 4, and 6), larger errors are seen when the actual number of sources is 8 and only the 'Relevant Set' size is varied. When the number of sources is 8, the errors found across the listening conditions all intersect. Note that providing a listener with a Relevant Set of 8 is no different than when 8 sources are presented and no Relevant Set is designated (the filled square and the open diamond). The difference between the curve representing data when the actual set size is varied and the curve depicting the data when only the Relevant Set size is varied can be attributed to differences in the signal-to-noise ratio across the two experimental conditions. However, note that the curve depicting performance as a function of the Relevant Set indicates that errors are much lower than would be expected if performance could be attributed to *only* signal-to-noise ratio. It appears that listeners are able to utilize the information provided by the presentation of the Relevant Set to attend to, and determine the location of, only a subset of the sources presented in this 8-source stimulus.

5.3 Conclusions

The results from Experiment 4 suggest that listeners are able to monitor and maintain an awareness of the spatial properties of multiple individual concurrent sounds in a complex environment. Their ability to both identify and localize sounds in Experiment 4a, and the way performance covaries across these two tasks, suggest that listeners are truly monitoring the individual sounds and their associated spatial properties, not just the overall spatial layout of an auditory scene. However, the results also indicate that listeners can benefit from a longer exposure to the auditory environment in order to judge characteristics of the sources in the environment. Whether this is due to the fact that listeners require time to

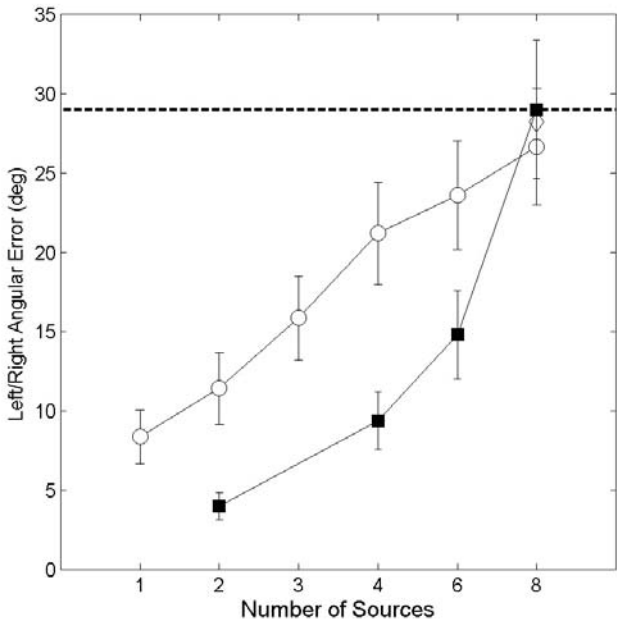

Figure 12: Left/Right localization errors plotted as a function of the number of sources when referring to the Relevant Set size (open circles) and the actual set size (filled squares). The open diamond is a condition in which no Relevant Set was designated. In both curves, error bars represent ± 1 standard error. The dashed line represents hypothetical error values if only the SNR determined performance.

serially attend to each sound, time to memorize multiple sounds, or time to establish a baseline set of interaural parameters against which to make their judgments, or in fact if all three aspects play a role is not clear, but it is clear that performance is quite good overall.

The results from Experiment 4b suggest that localization performance in these tasks was constrained by issues related to both signal-to-noise ratio and attention. The difference between the curve representing the actual set size and the curve for the relevant set size demonstrates the impact of signal-to-noise ratio on performance, but the contribution of attention to this task is also present. If only signal-to-noise ratio contributed to performance, localization errors would be independent of the size of the Relevant Set. However, given that performance actual varies as a function of the Relevant Set size suggests that attention plays an important role as well.

6.0 General Conclusions

The results from these studies suggest that both top-down and bottom-up processes contribute to a listeners ability to maintain awareness of the spatial properties of multiple concurrent sounds in a complex auditory environment. The results from Experiments 1-3 suggest that listeners can employ both analytic and synthetic listening strategies in order to accomplish these difficult tasks. That is, the results from these studies suggest that listeners can selectively attend to a specific sound, or divide their attention across multiple sounds. The results from Experiment 4 indicate that, even in these difficult environments, listeners have considerable knowledge of the sounds in an auditory scene and can maintain an awareness of the spatial properties of the sound sources with such scenes. Furthermore, these results seem to indicate that listeners can selectively attend to a subset of sounds presented in a larger set of sounds, and moreover can divide their attention across this subset of sounds in order to determine the location of the target sound. Whether this is achieved through serial or parallel processes is not clear, although it is likely that both processes play a role here. And although not explicitly examined, it is clear that memory could play in determining the constraints on performance, even though attempts were made to limit the role of memory in these tasks.

It is important to note that localization performance in these studies was more comparable to what would be expected from real-world experiences, and much better than would have been predicted by the results of spatial hearing experiments in other laboratories. It is believed that the use of environmental sounds contributes to this success. We believe that the ability to 'label' or 'name' a sound is a critical component in segregating, monitoring, and identifying multiple sounds in complex environments. Moreover, the novel techniques we have employed have enabled us to obtain access to the information a listener has in such environments in ways previously not obtainable. There is much to learn from this research effort, but it leaves many questions unanswered, and additional studies are already underway to address some of these questions.

This research was funded by a grant to the authors from the Air Force Office of Scientific Research (AFOSR).

References

Best, V., Carlile, S., Jin, C., and van Shaik, A. (2005). The role of high frequencies in speech localization. *Journal of the Acoustical Society of America*, 118:353–363.

Brungart, D. and Simpson, B. (2002). Within-channel and across-channel interference in the cocktail-party listening task. *Journal of the Acoustical Society of America*, 112:2985–2995.

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.

Drullman, R. and 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.

Eramudugolla, R., Irvine, D., McAnally, K., Martin, R., and Mattingley, J. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology*, 15:1108–1113.

Ghostwriters, D. D. (1998). The sound effects toolkit. CD.

Gilkey, R., Simpson, B., and Weisenberger, J. (1990). Masker fringe and binaural detection. *Journal of the Acoustical Society of America*, 88:1323–1332.

Gilkey, R. and Weisenberger, J. (1995). The sense of presence for the suddenly deafened adult. *Presence*, 4:357–363.

Gilkey, R. H. and Anderson, T. R. (1995). The accuracy of absolute localization judgments for speech stimuli. *Journal of Vestibular Research*, 5:487–497.

Good, M. D. and Gilkey, R. H. (1996). Sound localization in noise: The effect of signal-to-noise ratio. *Journal of the Acoustical Society of America*, 99:1109–1117.

Good, M. D., Gilkey, R. H., and Ball, J. M. (1997). The relation between detection in noise and localization in noise in the free field. In Gilkey, R. and Anderson, T., editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 349–376. Lawrence Erlbaum Associates, Mahwah, NJ.

Hirsh, I. J. (1948). The influence of interaural phase on interaural summation and inhibition. *Journal of the Acoustical Society of America*, 20:592–599.

Kistler, D. and Wightman, F. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91:1637–1647.

Langendijk, E. H. A., Wightman, F. L., and Kistler, D. J. (2001). Sound localization in the presence of one or two distracters. *Journal of the Acoustical Society of America*, 109:2123–2134.

Lorenzi, C., Gatehouse, S., and Lever, C. (1999). Sound localization in noise in normal-hearing listeners. *Journal of the Acoustical Society of America*, 105:1810–1820.

McFadden, D. (1966). Masking-level differences with continuous and with burst masking noise. *Journal of the Acoustical Society of America*, 40:1414–1419.

Miller, G. (1947). The masking of speech. *Psychological Bulletin*, 44:105–129.

Ramsdell, R. (1978). The psychology of the hard-of-hearing and the deafened adult. In Davis, H. and Silverman, S. R., editors, *Hearing and deafness, 4th Edition*, pages 499–510. Holt, Rinehart and Winston, New York.

Simpson, B. (2002). The localization of phantom auditory images. Master's thesis, Wright State University.

Simpson, B., Brungart, D., Gilkey, R., Iyer, N., and Hamil, J. (2006). Comparison of pre- and post-cueing in a multiple-source sound localization task. *Journal of the Acoustical Society of America*, 120:3081.

Wightman, F. and Kistler, D. (1989). Headphone simulation of free-field listening. ii: Psychophysical validation. *Journal of the Acoustical Society of America*, 85:868–878.

Wightman, F. L. and Kistler, D. (1997). Factors affecting the relative salience of sound localization cues. In Gilkey, R. and Anderson, T., editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 1–23. Lawrence Erlbaum Associates, Mahwah, NJ.