LOCALIZATION IN MULTIPLE SOURCE ENVIRONMENTS: LOCALIZING THE MISSING SOURCE

Brian D. Simpson[1], *Douglas S. Brungart*[1], *Robert H. Gilkey*[2], *Nandini Iyer*[1], *James T. Hamil*[3]

[1]Air Force Research Laboratory, WPAFB, OH
[2]Wright State University, Dayton, OH
[3]General Dynamics Advanced Information Systems, Dayton, OH
brian.simpson@wpafb.af.mil

ABSTRACT

Experience in real-world listening situations suggests that listeners, in general, have a great deal of spatial information about multiple concurrent sounds in an auditory scene. Despite this, laboratory data would suggest that listeners should operate quite poorly in such environments. This study employed environmental sounds that would naturally occur in real-world auditory environments and measured sound localization in auditory scenes containing 1, 2, 4, 6, or 8 concurrent sounds. The identifying feature of the target was that it was the only sound deleted from the multiple-source auditory scene at the end of an observation interval of a specific duration (2.5, 4.5, 6.5, or 8.5 sec). The results indicate that localization can be surprisingly good in complex auditory scenes. However, as an auditory scene becomes more complex, listeners appear to benefit from longer exposure to the scene in order to accurately judge the location of a change in the scene. [Work supported by AFOSR.]

[Keywords: Sound Localization, Multiple Sources]

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 [1, 2, 3]. 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.

Despite our 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.

In this paper we describe a study that employs this 'cueing by deletion' paradigm to examine a listener's ability to localize multiple sounds simultaneously. We varied 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.

2. GENERAL METHODS

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.

2.2. Apparatus

The experiment was conducted in the Auditory Localization Facility (ALF) in the Air Force Research Laboratory at Wright-Patterson Air Force Base (see Figure 1). This facility consists of a geodesic sphere (4.3m in diameter) with 277 Bose 11-cm, full-range loudspeakers mounted on its surface. The sphere is housed within an anechoic chamber, the walls, floor, and ceiling of which are covered in 1.1-m fiberglass wedges. For this study, only the 28 loudspeakers arranged along the horizontal plane of the ALF (i.e., the plane parallel to the ground that contains the interaural axis for an upright listener) were utilized. These loudspeakers are spaced approximately every 15° on the horizontal plane. In addition, loudspeakers located at positions directly in front of, behind, and to the sides of the listener were included. Mounted on the front of each loudspeaker is a square cluster of four light-emitting diodes (LEDs).

An Intersense IS-900 ultrasonic headtracker, attached to a headband worn by the listener, was used to determine the orientation of the listener's head. This information was used to enforce stationary head orientation throughout the stimulus presentation interval,

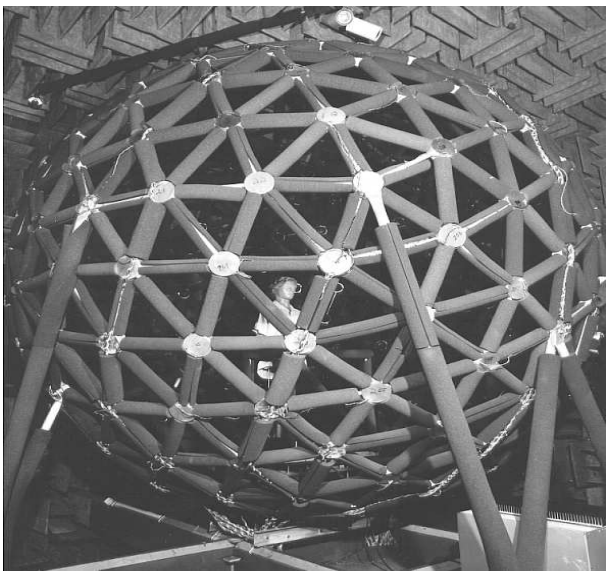

Figure 1: *The Auditory Localization Facility at Wright-Patterson Air Force Base. See text for details.*

and also as a localization response mechanism. Specifically, orientation information from the headtracker was used to activate the LED cluster directly in front of the listener such that as the listener's head orientation changed, so too did the location of the activated LEDs (a 'head-slaved LED cursor'). A button on a handheld response device was depressed when the desired LED cluster was activated (i.e., when the listener had oriented to the desired response location). Individual audio signals were routed from a control computer to a Mark of the Unicorn digital-to-analog converter (MOTU 24 I/O), then through a bank of amplifiers (Crown Model CL1), and finally directed to the appropriate loudspeaker through a custom-built loudspeaker switching system (Winntech).

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 [4]. 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°. 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.

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° 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° 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° 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.

3. RESULTS

3.1. Experiment 1

For analysis purposes, the azimuthal localization errors were decomposed into a left/right component and front/back component [5]. 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.

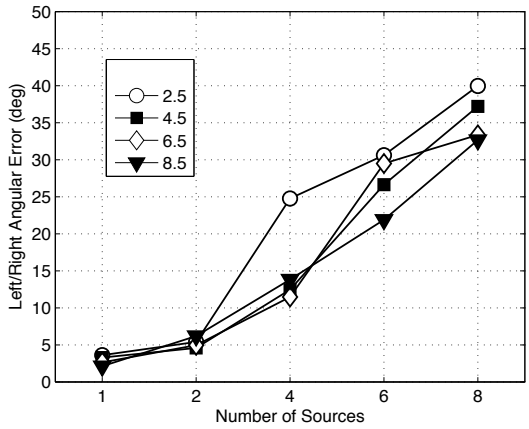

Figure 2: *Left/Right localization errors, averaged across all listeners, plotted as a function of the number of sources for each duration of observation interval.*

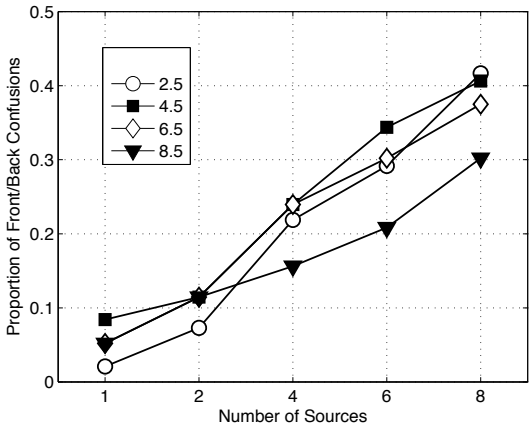

Figure 3: *Proportion of front/back confusions, averaged across all listeners, plotted as a function of the number of sources for each duration of observation interval.*

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 2, 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 2, 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° of the actual location, suggesting that these individual sounds were sufficiently broadband to support good left/right localization.

Overall, the data from Figure 2 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° 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 3, 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.

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.

The results from Experiment 2 are shown in Figure 4. 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 sig-

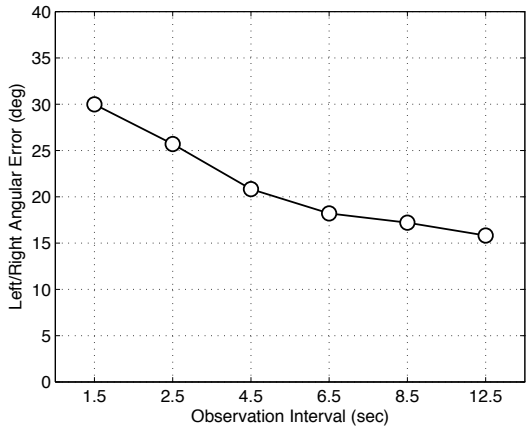

Figure 4: *Left/right localization errors, averaged across listeners, plotted as a function of the duration of the observation interval for 6 simultaneous sources.*

nificant 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° vs 30°).

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.

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 [6] 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° larger than those in the current study under comparable conditions. In part, the larger errors found in [6] 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° 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 [7], which is presumably a simpler task than localization. This recent study [7] 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 [6] or noise [7]. 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 [8], suggesting that this background provides a context against which to perceive these changes. Moreover, in both [6] and [7], 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 [9].

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.

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.

6. REFERENCES

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

[2] E. H. A. Langendijk, F. L. Wightman, and D. J. Kistler, "Sound localization in the presence of one or two distracters," *Journal of the Acoustical Society of America*, vol. 109, pp. 2123–2134, 2001.

[3] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 105, pp. 1810–1820, 1999.

[4] Dave Dworkin's Ghostwriters, "The sound effects toolkit," CD, 1998.

[5] D.J. Kistler and F.L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America*, vol. 91, pp. 1637–1647, 1992.

[6] B.D. Simpson, D.S. Brungart, R.H. Gilkey, N. Iyer, and J.T. Hamil, "Comparison of pre- and post-cueing in a multiple-source sound localization task," *Journal of the Acoustical Society of America*, vol. 120, pp. 3081, 2006.

[7] R. Eramudugolla, D.R.F. Irvine, K.I. McAnally, R.L. Martin, and J.B. Mattingley, "Directed attention eliminates 'change deafness' in complex auditory scenes," *Current Biology*, vol. 15, pp. 1108–1113, 2005.

[8] R.H. Gilkey, B.D. Simpson, and J.M. Weisenberger, "Masker fringe and binaural detection," *Journal of the Acoustical Society of America*, vol. 88, pp. 1323–1332, 1990.

[9] D. McFadden, "Masking-level differences with continuous and with burst masking noise," *Journal of the Acoustical Society of America*, vol. 40, pp. 1414–1419, 1966.