THE IMPACT OF HEADTRACKING ON INTELLIGIBILITY IN A MULTITALKER DISPLAY

Douglas S. Brungart[1], Brian D. Simpson[1], Alexander J. Kordik[2], and Richard L. McKinley[1]
[1]Air Force Research Laboratory and [2]Sytronics, Inc.
Wright-Patterson AFB, Ohio

Although overall performance in multichannel speech monitoring tasks is known to improve substantially when the apparent locations of the competing talkers are spatially separated with a virtual audio display, the impact that real-time interactive headtracking has on these tasks is not yet well understood. In this experiment, listeners were asked to monitor four simultaneous spatially-separated speech signals and to identify the contents of a target phrase that was addressed to a pre-assigned call sign. This task was performed with and without interactive headtracking, with wide (60°) and narrow (20°) spatial separations between the competing talkers, and with four different probabilistic models for changes in the location of the target talker. The results indicate that interactive headtracking cues have a significant effect on multitalker listening performance only in situations where the target talker tends to remain in the same location for more than one consecutive stimulus presentation, and that the addition of interactive headtracking cues in these situations can either improve or degrade overall performance depending on the size of the angular separations between the competing talkers in the virtual environment.

INTRODUCTION

In command, control, and communication tasks that require listeners to monitor multiple simultaneous channels of speech, substantial performance improvements can be obtained with virtual audio displays that spatially separate the apparent locations of the competing talkers (Begault, 1999; Brungart, Ericson, & Simpson, 2002; Crispien & Ehrenberg, 1995; Drullman & Bronkhorst, 2000; Ericson & McKinley, 1997). There is also evidence that many of the intelligibility benefits afforded by virtual sound source separation can be achieved with a much lower level of audio display fidelity than would typically be required to provide a listener with robust information about the locations of virtual sounds. For example, the results of experiments examining multitalker speech intelligibility in virtual audio displays indicate that overall performance is roughly the same with non-individualized head-related transfer functions (HRTFs) as it is with individualized HRTFs (Drullman & Bronkhorst, 2000). This is in direct contrast to studies that have shown that virtual sound source *localization* is much more accurate when individualized HRTFs are used, especially in the front-back and up-down dimensions[1] (Wenzel, Arruda, Kistler, & Wightman, 1993). Large speech intelligibility advantages have also been demonstrated with non-interactive virtual audio displays that lack the capability to update the relative locations of the sound source in response to the exploratory motions of the listener's head, despite the critical role that head coupling is known to play in producing well-externalized virtual sound images and in resolving front-back confusions in sound localization (Wightman & Kistler, 1999). However, the positive results achieved in multitalker listening studies with non-head-coupled virtual audio displays can only provide evidence that headtracking is not *necessary* to achieve a benefit from the spatial separation of competing talkers; the extent to which there might be an *additional* benefit from the use of a headtracked virtual audio display remains an open question. In this paper, we describe an experiment that examined the effects of real-time headtracking in a task that required listeners to monitor four simultaneous spatially-separated channels of speech.

[1]Note that this discrepancy is likely due in part to the fact that talkers in multitalker listening studies are usually separated in the left-right dimension where individual cues HRTF cues are relatively unimportant for localization and in part to the fact that individual HRTF differences are small in the frequency range below 3500 Hz where most speech information resides.

METHODS

Participants: A total of 11 paid volunteer listeners (5 male and 6 female) participated in the experiment. All had normal hearing ($<$ 15 dB HL from 500 Hz to 8 kHz), and their ages ranged from 18 to 50 years. None had any experience or practice in a multitalker listening task with headtracking, but all had previously participated in experiments that involved multitalker listening with the same speech materials used in this experiment.

Speech Materials: The experiment was conducted with the publicly-available Coordinate Response Measure (CRM) speech corpus (Bolia, Nelson, Ericson, & Simpson, 2000). This corpus consists of 2048 phrases of the form "Ready, (call sign), go to (color) (number) now," comprised of all combinations of four colors ("red," "blue," "green," or "white"), eight numbers (1-8), eight call signs ("Baron," "Charlie," "Ringo," "Eagle," "Arrow," "Hopper," "Tiger," and "Laker") spoken by four male and four female talkers. On average, the phrases in the corpus are approximately 1.75 s long. Note that the phrases in the speech corpus have been hand aligned to synchronize the start of the introductory word "ready," and that the entire corpus has been low-pass filtered at 8 kHz.

Audio Spatialization: In each trial of the experiment, the listeners were presented with four simultaneous spatially-separated phrases from the CRM corpus. The sentences were

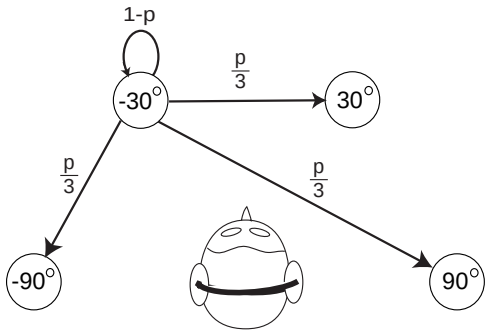

Figure 1. Transition diagram for the $\Delta = 60°$ condition of the experiment, where the talker at $-30°$ is currently designated as the target talker. On the next trial, the control computer will randomly select a different talker to be the target talker with probability p, and keep the same talker designated as the target talker with probability $1 - p$.

generated on four different output channels of an 8-channel D/A system (TDT DA3-8), and these channels were then fed into four inputs of a 6-channel real-time spatial audio display (Veridian 3-DVALS System II) that was loaded with HRTFs collected every one degree in azimuth in the horizontal plane with a KEMAR acoustic manikin at a source distance of 50 cm (Brungart & Rabinowitz, 1999). The 3-D VALS system was also connected to a gyroscopic headtracking device (Intersense IS-300) that was attached to the listener's headphones. In the experimental conditions with headtracking, the HRTFs associated with the four input channels were updated in real time to compensate for the listener's head movements. In the conditions without headtracking, the HRTFs were not updated to compensate for head motion, but the listener's head movements were still recorded at roughly 200 ms intervals throughout the duration of the speech stimuli.

Spatial Configurations: Two different spatial configurations were tested in the experiment. In the $\Delta = 60°$ condition, the four competing talkers were located at $-90°$, $-30°$, $30°$, and $90°$ in azimuth, as illustrated in Figure 1. In the $\Delta = 20°$ condition, the talkers were located at $-30°$, $-10°$, $10°$ and $30°$ in azimuth. At the start of each block of trials, each of the four male talkers in the CRM corpus was randomly assigned to one of the four possible source locations. These talkers remained fixed at these locations throughout that 60-trial block, and the data collection was balanced so that each listener heard each of the four talkers at each of the four starting locations in every combination of conditions tested in the experiment.

Transition Probability: In real-world listening tasks, listeners who are monitoring multiple sources of information are constantly required to shift their attention to the source that they perceive to be the most interesting or relevant at that particular moment in time. Because the frequency of these required attention shifts can vary substantially across different kinds of listening tasks, it is useful to measure the performance of multitalker speech display systems as a function of the frequency of these target talker transitions. In this experiment, the frequency of these transitions was controlled with the discrete probability model illustrated in Figure 1. At the start of each block of trials, one of the four spatially-separated talkers in that condition was selected to serve as the initial target talker. This talker was always distinguished from the other talkers in the stimulus by the use of the call sign "Baron" in the CRM carrier phrase. Then, at the end of that trial and each subsequent trial, the identity and location of the target talker in the next trial was selected according to the transition diagram shown in Figure 1. In other words, the target phrase was randomly switched to one of the other locations with probability p and remained at the same location with probability $1 - p$. A Four different values of p were tested in the experiment: 0.125, 0.25, 0.5, and 1.0.

Procedure: The listeners participated in the experiment while seated in front of the CRT of a Pentium-III based control computer in a quiet listening room. Prior to each block of trials in the experiment, the listeners were asked to boresight the headtracker by fixating on a point in the middle of the CRT and pressing a key. Then they were given instructions that indicated whether the stimuli would be headtracked in that particular block of trials. In blocks with headtracking, they were informed that they "might benefit from head movement". In blocks without headtracking, they were told that "head motion would have no effect."

Once these instructions were acknowledged with a key press, data collection began immediately. In each trial, the listener heard a stimulus containing four simultaneous CRM phrases: the target phrase, which was identified by the presence of the call sign "Baron," and three masking phrases, which contained call signs other than "Baron." The phrases were selected randomly from the corpus with the restriction that no two phrases ever contained the same call signs and none of the masking phrases ever contained the same color or number as the target phrase. The task was to listen for the target phrase addressed to the call sign "Baron" and respond by using the mouse to select the color and number combination contained in that phrase from an array of colored digits displayed on the CRT of the control computer. This effectively forced the listeners to return their gaze to the CRT between each pair of consecutive trials, which is consistent with the wide variety of operational tasks that require listeners who are monitoring multiple communications channels to simultaneously interact with a conventional CRT-based computer interface.

Experiment Design: The data were collected in a 3-factor within-subjects repeated measures design in which each subject participated in four 60-trial blocks with each of the 16 possible factorial combinations of two spatial configurations ($\Delta = 60°$ or $20°$), two headtracking conditions (tracked or non-tracked), and four transition probabilities (0.125, 0.25, 0.5, or 1.0). Thus, over the course of several weeks each of the 11 listeners participated in a total of 64 blocks containing 3,840 trials, for a total of 42,240 trials collected in the experiment.

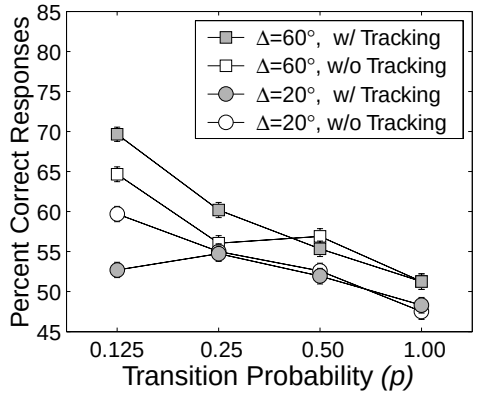

Figure 2. Percentage of correct color and number identifications as a function of the transition probability *p* for each spatial configuration and each headtracking condition of the experiment. The error bars represent ± 1 standard error around each data point.

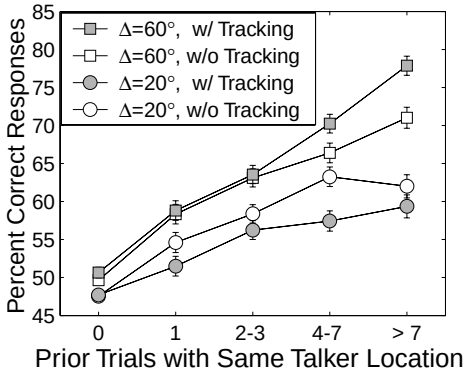

Figure 3. Percentage of correct color and number identifications as a function of the number of previous trials the target talker has remained at the same source location. The geometric bin spacing was selected to try to equalize the number of trials in each bin: each symbol for 0 prior trials represents roughly 5100 data points, each symbol for 1, 2-3, and 4-7 prior trials represents roughly 1400-1600 data points, and each symbol for > 7 prior trials represents roughly 1100 data points. The error bars represent ± 1 standard error around each mean value.

RESULTS AND DISCUSSION

The overall results of the experiment are summarized in Figure 2, which shows the percentage of correct color and number identifications with and without headtracking for each combination of spatial configuration and transition probability tested in the experiment. These same data were also analyzed with a 3-factor within-subject repeated-measures ANOVA, where the arcsine-transformed percentage of correct responses for each 60-trial block was counted as a single replication of each condition. In general, overall performance was better in the $\Delta = 60^\circ$ condition than in the $\Delta = 20^\circ$ (squares versus circles in the figure, main effect $F(1,10)=15.93$, $p<0.005$), and in the low-transition-probability conditions than in the high-transition-probability conditions (moving from left to right in the figure, main effect $F(3,30)= 25.56$, $p<0.001$).

However, the addition of headtracking did not have consistent effects across the different conditions of the experiment. The main effect of headtracking was not significant ($F(1,10)=0.004$, $p>0.9$). There was, however, a significant 3-way interaction between headtracking, spatial configuration, and transition probability ($F(3,30)=5.14$, $p< 0.01$). When the transition probability *p* was 0.5 or 1.0, headtracking had almost no effect on performance. When *p* was 0.25 or 0.125 in the $\Delta = 60^\circ$ condition, headtracking improved performance by approximately 5 percentage points. And when *p* was 0.125 in the $\Delta = 20^\circ$ condition, headtracking actually *degraded* performance by approximately 8 percentage points.

These data indicate that real-time headtracking has a much greater effect on performance in conditions with low transition probabilities, where the target talker tends to remain in the same location for multiple consecutive trials, than at high transition probabilities, where the target talker is rarely in the same place for more than a few trials at a time. This suggests that a listener's ability to make use of real-time headtracking cues varies with the number of trials that target talker remains in the same location. This effect is illustrated in Figure 3, which plots the results from each condition of the experiment as a function of the number of previous trials in which the target talker has remained at the same location. Thus, the leftmost point (0) indicates trials where the target talker has just moved, and the rightmost point (> 7) indicates trials where the talker has remained in the same place for 8 or more consecutive trials[2]. The results from the $\Delta = 60^\circ$ condition show that the advantages of headtracking with widely-spaced sources take a number of trials to build up (four or more trials in the CRM task), but that once they do come into play they produce a substantial improvement in overall performance. However, the results from the $\Delta = 20^\circ$ condition show that the disadvantages of interactive headtracking with narrowly-spaced sound sources occur very quickly (after only one trial). Taken together, the results shown in Figures 2 and 3 indicate that real-time headtracking improves performance only in tasks where the spatial locations of the virtual sources are widely distributed *and* the target talker location changes relatively infrequently, and suggest that real-time headtracking should be avoided altogether in cases where the competing speech channels are narrowly spaced.

CONCLUSIONS

The results of this experiment lead to somewhat mixed conclusions about the effect that real-time headtracking has on the performance of a multitalker speech display. Prior to conducting the experiment, our expectation was that performance would always be at least as good with headtracking as it would

[2]Note that this measure is highly correlated with transition probability, in that the conditions with low transition probability will have a much greater incidence of trials with a large number of prior trials in the same location, and the condition with *p*=1 will have only trials with no prior trials in the same location.

be without headtracking, if for no other reason then simply because the listener would always have the option of not moving the head. However, these results suggest that real-time headtracking can actually hurt performance when the spatially-separated talkers are located relatively close to one another. A plausible explanation for this result is related to the changes in spatial acuity that are known to occur at different sound source locations around the head. In general, listeners are much more sensitive to changes in the azimuth of sound sources directly in front of them than they are to changes in the azimuth of sound sources off to the sides. Mills, for example, showed that listeners are 6-10 times more sensitive to changes in the azimuth of sound sources near 0° than they are to changes in the azimuth of sounds near $\pm\ 90^\circ$ (Mills, 1958). Consequently, one might expect two sound sources separated by 20° to be easier to segregate when they are located at $\pm 10^\circ$ in azimuth than when they are located at 10° and 30° or 20° and 40°. By allowing listeners to move their heads in the $\Delta = 20^\circ$ condition of this experiment, it is possible that we were encouraging them to move to an orientation off the midline that failed to optimize the effective spatial separation of the different talkers in that condition. This might account for the relatively poor performance that was observed when headtracking was enabled in that condition. While the obvious solution to this problem in the four-talker case is to use a wider talker separation, it should be noted that this issue might be more difficult to address in speech display systems that attempt to spatially separate more than five simultaneous channels of speech.

In the $\Delta = 60^\circ$ condition, where the talkers were spaced relatively far apart, real-time headtracking did provide a quantifiable performance benefit. However, this benefit was limited to cases where the location of the talker changed infrequently, and even in those cases it produced only a modest improvement in performance (5-8 percentage points). This contrasts sharply with the much larger improvement in performance that occurs when a multitalker speech display is changed from a diotic or nonspatialized system to a binaural system that spatially separates the apparent locations of the competing talkers without interactive headtracking. For example, in the 4-talker CRM task tested in this experiment, previous experiments have shown that listeners are able to correctly identify the color and number in the target phrase only 25% of the time with a non-spatialized display (Brungart et al., 2002). Thus, in the 4-talker CRM task, the percentage of correct responses improves from roughly 25% to roughly 50% when a diotic display is replaced with a non-headtracked spatialized display where the target talker moves frequently ($p >= 0.5$) and to roughly 65% with a non-headtracked spatialized display where the target talker moves infrequently ($p = 0.125$), but only by an additional 5% when real-time headtracking is added to a spatialized display with an infrequently moving target talker. Thus, while it is true that headtracking can provide some improvement in performance in a spatialized multitalker speech display, this performance improvement is small relative to the large improvement that can be obtained by adding non-headtracked spatial processing to a monaural or diotic display system. Since real-time headtracking remains one of the most expensive capabilities to add to a virtual audio display, this result may argue against the use of headtracking in display systems that are designed exclusively for use in multichannel communications tasks.

At the same time, it should be noted that situations may occur where a multitalker audio display is needed in an environment that is already configured to collect headtracking data, either for an audio display intended to convey sound localization information or for another type of display, such as a visual head-mounted display (HMD), that also requires real-time head position information. In these situations, the intelligibility advantages afforded by interactive headtracking could easily justify the relatively modest additional cost required to interactively head-couple the spatially-separated communications channels in the system. Headtracking might also provide other advantages in these environments, such as increased situational awareness, greater intelligibility in noise, and reduced workload. However, the results from the $\Delta = 20^\circ$ condition of this experiment suggest that care should be taken to ensure that the competing speech channels are not placed too close together, or there is a risk that this headtracking could degrade, rather than enhance, the performance of the system.

References

Begault, D. R. (1999). Virtual Acoustic Displays for Teleconferencing: Intelligibility Advantage for 'Telephone-Grade' Audio. *J. of the Aud. Eng. Soc.*, *47*, 824–828.

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

Brungart, D., Ericson, M., & Simpson, B. (2002). Design considerations for improving the effectiveness of multitalker speech displays. In *Proceedings of the International Conference on Auditory Display, Kyoto, Japan, July 2-5, 2002*, pp. 169–174.

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

Crispien, K. & Ehrenberg, T. (1995). Evaluation of the 'Cocktail Party Effect' for Multiple Speech Stimuli within a Spatial Audio Display. *J. of the Aud. Eng. Soc.*, *43*, 932–940.

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

Ericson, M. & McKinley, R. (1997). The intelligibility of multiple talkers spatially separated in noise. *In Binaural and Spatial Hearing in Real and Virtual Environments*. Edited by R.H. Gilkey and T.R. Anderson (Erlbaum, Hillsdale N.J.), pp. 701-724.

Mills, A. (1958). On the minimum audible angle. *J. Acoust. Soc. Am.*, *30*, 237–246.

Wenzel, E., Arruda, M., Kistler, D., & Wightman, F. (1993). Localization using non-individualized head-related transfer functions. *J. Acoust. Soc. Am.*, *94*, 111–123.

Wightman, F. & Kistler, D. (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.*, *105*, 2841–2853.