RESPONSE TECHNIQUES AND AUDITORY LOCALIZATION ACCURACY

Nandini Iyer, Eric R. Thompson, Brian D. Simpson

Air Force Research Laboratory, 711 Human Performance Wing
2610 Seventh St., B441 Wright-Patterson AFB, OH 45433 USA
nandini.iyer.2@us.af.mil

ABSTRACT

Auditory cues, when coupled with visual objects, have lead to reduced response times in visual search tasks, suggesting that adding auditory information can potentially aid Air Force operators in complex scenarios. These benefits are substantial when the spatial transformations that one has to make are relatively simple i.e., mapping a 3-D auditory space to a 3-D visual scene. The current study focused on listeners' abilities to map sound surrounding a listener to a 2-D visual space, by measuring performance in localization tasks that required the following responses: 1) Head pointing: turn and face a loudspeaker from where a sound emanated, 2) Tablet: point to an icon representing a loudspeaker displayed in an array on a 2-D GUI or, 3) Hybrid: turn and face the loudspeaker from where a sound emanated and them indicate that location on a 2-D GUI. Results indicated that listeners' localization errors were small when the response modality was head-pointing, and localization errors roughly doubled when they were asked to make a complex transformation of auditory-visual space (i.e., while using a hybrid response); surprisingly, the hybrid response technique reduced errors compared to the tablet response conditions. These results have large implications for the design of auditory displays that require listeners to make complex, non-intuitive transformations of auditory-visual space.

1. INTRODUCTION

In the Air Force, operators are routinely required to make complex decisions with an incredible barrage of information, often under severe cognitive load (multitasking, multiple sources to monitor, interoperability issues, etc.). Most information is presented using visual interfaces, and due to increasing complexity of operations, there is a high likelihood that operators might miss critical information if these events occur outside of the locus of visual attention. In these complex operations, the consequences of missing critical information could be greatly reduced by introducing multimodal displays and presenting some of the information through the auditory modality. The auditory modality has the advantage of responding to sounds arriving from anywhere in the environment; thus, while the spatial resolution of the auditory system is coarse relative to visual spatial resolution, its coverage is greater (360 degrees), reducing the possibility that events occurring outside the field of view will go undetected. Auditory cues can also effectively increase awareness of one's surroundings, convey a variety of complex information without taxing the visual system, increase the sense of presence in immersive environments, cue visual attention, and facilitate cross-modal enhancements.

One of the early obstacles to using auditory cues in operational environments was the feasibility and cost of recreating auditory space over headphones. Using signal-processing techniques, it is now relatively easy to generate stereophonic signals under headphones that recreate the spatial cues available in the real-world; in fact, when these cues are rendered appropriately, it is often difficult to distinguish between sounds presented in the free-field over loudspeakers from those presented virtually over headphones [1]. When coupled with a head-tracker, realistic acoustic environments that respond naturally to dynamic source and head motion can be rendered in 3-dimensional (3-D) space around a person's head. Such displays are not only effective and compelling, but are now sufficiently mature for integration into systems employed in operational environments.

Several studies have now shown that auditory cues can speed responses to targets in a visual search task, i.e., when auditory cues are co-located with visual targets, search times for a visual target are reduced significantly compared to the same search conducted without an auditory cue. Further, some of these studies demonstrated that an auditory cue reduced the deleterious effects on response time that typically occur with increases in visual scene complexity [2, 3]. In all of these experiments, accuracy was fairly high by design and only response time differences were measured and reported. While response time is an important metric in real-world operational environments, research studies are also interested in the question of localization accuracy. In determining localization accuracy, a number of different response techniques have been used; a relatively natural response method requires listeners to turn to the stimulus location and localization responses are obtained by tracking the listener's head in space (head-pointing, finger-pointing or nose-pointing response methods) [4]. Other natural response methods have included using a pistol-like device to "shoot" at the target location [5]. While head-, finger- or nose-pointing responses are most accurate [6], by and large, these other response methods yield comparably similar localization responses; in part, it is due to the fact that these localization responses do not require any mental transformations of the target location and the listeners can use their own anatomical references to localize a sound.

In contrast to more direct or natural localization responses, indirect localization response methods have also been used, such as a verbal reporting technique [7], God's eye localization pointing (GELP) [6], or the large- and small-head response techniques [8]. In verbal response methods, listeners had to be trained to state the perceived azimuth and elevation on sources that were then transcribed by an experimenter. In the GELP method, listeners were

DOI: 10.21785/icad2016.032

Figure 1: Picture of the facility with the circular speaker array. A listener is shown seated in the center of the array with the tablet and head tracker.

required to make their responses of perceived location on a 22 cm plastic sphere using an electromagnetic sensor. In the large- and small-head response techniques, the sphere in GELP was replaced by an anatomically correct styrofoam head (large or small); the head was placed in front of the subject facing their right side for easy accessibility. All of these procedures were less accurate than direct localization responses, suggesting that some response techniques yield better localization accuracy than others. There are several reasons for decreased accuracy of localization responses with indirect methods; for example, the verbal response technique required correct interpretation and entry by an experimenter which could be a source of error. The remaining three response techniques require listeners to remember target sources relative to their own head and then transform that response onto another representation of possible source location (either a sphere or a styrofoam head). In summary, indirect localization responses are presumably less intuitive than direct pointing techniques.

In future AF operations, we can foresee several areas where operators can utilize natural, intuitive localization judgments to acquire a visual target or avoid a threat; for example, head- mounted displays (HMDs) can be head-tracked to display coupled visual and auditory information allowing operators to access data that are tied to line-of-sight. However, in order for audio-aided cueing to be effective in operational environments, operators might have to make more complex transformations of an auditory cue to an associated visual space; for example, in intelligence, surveillance and reconnaissance (ISR) missions, operators are tasked with trying to find, fix and track a target when presented with a God's eye view of visual targets on the ground. In such scenarios, it is not unreasonable to expect that an audio cue might assist operators to locate a target and track a moving target. However, it is not clear if operators could make the spatial transformations required to best utilize such a cue. That is, can an observer benefit from a 3-D spatial audio cue generated and presented virtually to reduce the acquisition time of a visual object presented on a 2-D display located on a monitor/screen in front of the operator. The current experiment was designed to assess if localization accuracy and response time would vary as a function of the response technique employed in the task. Three response techniques were evaluated: head-pointing, tablet response (with possible target locations displayed using a GUI) and a hybrid method that incorporated head-turning to localize the sound on a tablet. The first technique is very intuitive and

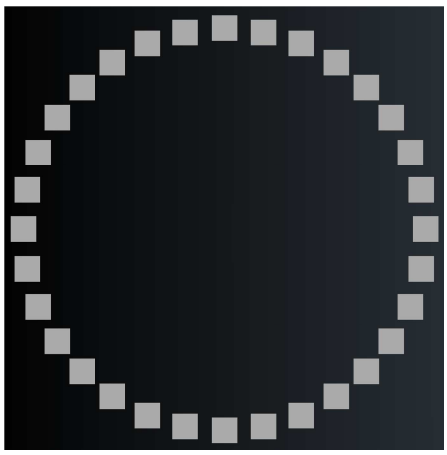

Figure 2: Screenshot of the tablet Graphical User Interface (GUI) used with the tablet and hybrid methods.

natural where a listener has to turn his/her head towards the direction of a sound. The second technique requires listeners to perform a more complex transformation of space; he/she has to associate a sound with an object representing the loudspeakers displayed on a tablet. The third technique was employed to evaluate whether or not turning and acquiring a source might facilitate accuracy with a tablet response.

Another factor that is important while evaluating response techniques is the actual stimulus itself. It is known that high frequency energy (above 8 kHz) is important for accurate auditory localization of non-speech signals [9, 10, 11, 12]. The few studies that have examined speech localization in the horizontal plane using single words presented found no significant differences in localization accuracy with speech and non-speech broadband stimuli; however, they reported an increase frontback confusions with speech stimuli [6]. In the current study, both non-speech broadband stimuli as well as speech were used as stimuli to evaluate whether the two response techniques might have differential effects on the two types of stimuli.

2. METHOD

2.1. Participants

Eleven listeners (six female) participated in the experiment. All had normal audiometric thresholds ($<$20 dB HL at octave frequencies between 250 and 8000 Hz) and normal or corrected-to-normal vision, and all had prior experience in sound localization studies. They were all paid to participate in the study, and all provided informed consent under a protocol approved by the Air Force Research Lab, 711th HPW Institutional Review Board.

2.2. Stimuli

On each trial, the listeners heard either a burst of noise, or a single word, which were both presented at a sampling rate of 44.1 kHz. The noise bursts had a pink spectrum between 100 Hz and 20 kHz, and a duration of 250 ms. The words were from recordings made in our lab of the PB-50 word lists [13]. The word lists were recorded

with a sampling rate of 44.1 kHz by twelve talkers (six female), none of whom were participants in this experiment. The recordings were post-processed to normalize the duration of each utterance to 500 ms using the Praat software [14].

2.3. Equipment

The experiment was controlled by a PC running the Windows 7 operating system and Matlab (the Mathworks, R2013a). The stimuli were generated in Matlab, and were presented through a RME RayDAT interface, RME M-32DA converter, eight Crown CTs 4200 four-channel amplifiers, and 32 Orb Audio loudspeakers in an evenly spaced circular array, approximately 11.25 degrees apart in azimuth (see Fig. 1). The listener was seated at the center of the array with their ears on the same plane as the loudspeaker array. An Optitrack V120:Trio was used with a custom array of reflecting balls mounted on a headband for tracking head motion. A cluster of four LEDs was mounted in front of and centered on each loudspeaker. The LED array was controlled by PacLED64 boards (Ultimarc). On head-tracked response trials, input was made using a Nintendo Wii remote, with communication to Matlab provided by WiiLab [15]. The tablet interface was a custom app running on a Google Nexus 9 tablet.

2.4. Procedure

Within a block of 32 trials, the stimulus condition (noise or speech) was held constant, and the source location was pseudo-randomly selected so that each listener completed 15 trials per source location per response method. On each trial, the listeners oriented toward a reference location (defined as 0 degree azimuth), heard a stimulus, made a response according to the response input method for that block, and received feedback for the correct location of the source. The details for each response method follow. Each listener completed all of the blocks for one response method before continuing to the next method. The head tracking and tablet response methods were completed first and with six listeners completing the head tracking method first and the remaining completing the tablet response first. Both groups (tablet first or head tracking first) completed the hybrid response as the third condition.

2.4.1. Head pointing method

At the start of each trial, the listener was required to orient towards a reference loudspeaker. The LED cluster closest to their head orientation would illuminate as their head turned. When the LED cluster at the reference location was illuminated, they pressed a button on the Wii remote to start the stimulus presentation. After the end of the stimulus, the head-slaved LED cursor would again illuminate, and the listeners would turn to look in the perceived direction of the sound. When the LED cluster was illuminated on the desired response location, they again pressed a button on the Wii remote to enter their response.

2.4.2. Tablet method

The listeners remained oriented toward the reference loudspeaker for the whole experiment. The tablet showed a circular array of 32 buttons that corresponded to the 32 loudspeaker locations (see Fig. 2). The reference location was at the top of the tablet display. In each trial, they heard the stimulus, and then indicated on the tablet the location from which they perceived the sound to originate. The correct source location was indicated after their response by flashing the background color of the corresponding button location on the GUI.

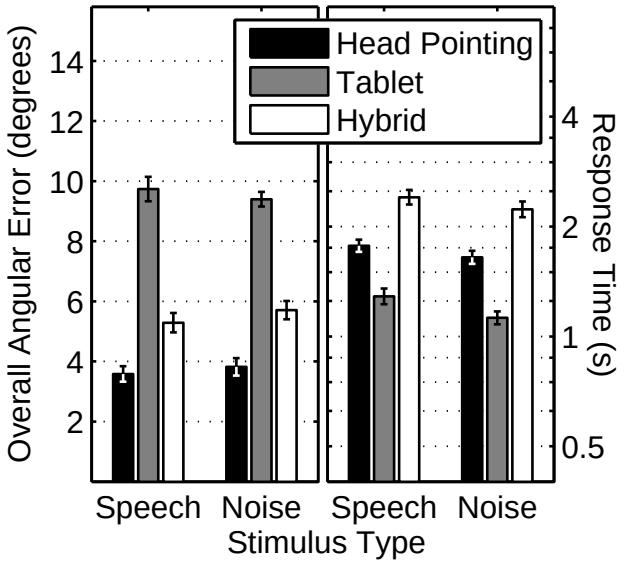

Figure 3: Average unsigned localization error (left panel) and mean response time (right panel: note logarithmic response times) plotted as a function of the two stimuli types, speech and noise. The parameters in the figures represent the three different response techniques used: head pointing (black bars), tablet response (gray bars) and hybrid response (head-pointing then responding using the tablet: white bars). Error bars represent standard error for within-subject measures.

2.4.3. Hybrid method

The hybrid response method contained elements of the head tracking and tablet methods. The listeners had to orient toward the reference loudspeaker before each trial and press a button on the tablet to begin stimulus presentation. After the stimulus ended, they were instructed to turn their head and look in the perceived direction of the sound source. After identifying the sound source location, they were instructed to return to the reference position and select the response on the tablet (the tablet display did not rotate with the head orientation). As with the tablet method, correct location feedback was provided by flashing the background color of the corresponding button.

3. RESULTS AND DISCUSSION

The data from the experiment are plotted in Figure 3; the left panel depicts average angular error as a function of the stimuli used in the experiment, for the three types of response techniques. As is apparent in the figure, there appears to be no difference in localization accuracy for the two different types of stimuli (Speech and Noise) across response techniques used. The response techniques do, however, influence localization accuracy. Specifically, local-

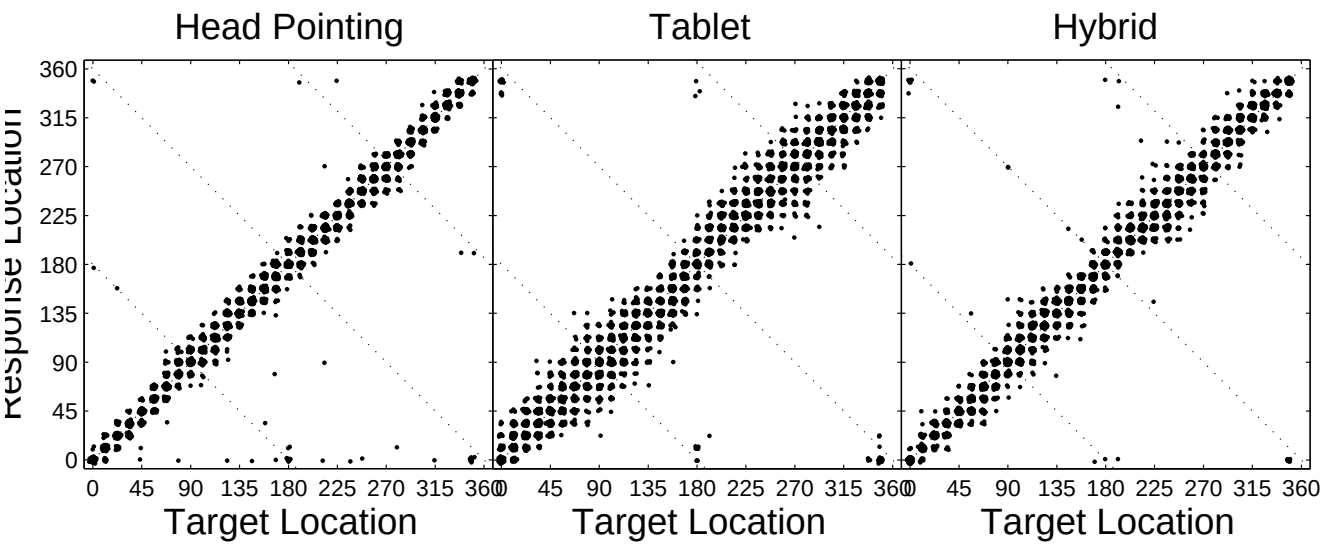

Figure 4: Scatter plot of response location vs. source location for the three response methods. Each data point in the plot is jittered slightly so that more frequently used target/response combinations appear with a larger black patch. The dashed lines with a negative slope in each plot indicate where front-back and left-right reversals would lie.

ization accuracy was best (about 4 deg.) for the head-pointing response technique and increased by at least a factor of two (to about 9.7 deg.) with the tablet response technique. However, when they first turned to look at a loudspeaker before responding on the tablet, errors were smaller when compared to the tablet response technique. The right panel in Figure 3 depicts the geometric mean response times for the three different response techniques in the experiment. As is evident from the figure, listeners were fastest using the tablet response (approx. 1.5 sec on average), slowest using the hybrid technique (approx. 2.3 sec) and took approximately 1.75 sec on average to respond using head-pointing. This was true for both Speech stimuli and Noise stimuli. It is perhaps not surprising that response times were longer using the head-pointing technique, since listeners had to rotate in their chairs and look at a loudspeaker, whereas in the tablet response condition, they responded by clicking on a GUI display of the loudspeaker array. Interestingly, response times were faster in the hybrid condition than the simple sum of response times for the head-pointing and tablet response techniques, especially since they had to reorient towards the reference speaker before they responded; it is possible that the reduction in response time occurred due to reduced transformations listeners made in the hybrid condition compared to the tablet condition. Another possibility is that listeners were generally more efficient at associating a visual target to a source, presumably because the head-pointing response relies on explicit memory of the source's visual location, and the mental transformation onto a tablet response was easier in this condition.

Figure 4 depicts performance for the three response techniques as a scatter plot, where target locations are plotted on the abscissa and response locations are plotted along the ordinate. In these plots, correct responses would fall on the diagonal from lower-left to upper-right. As seen in the figure, head pointing response techniques resulted in a tight distribution with most of the responses falling along the diagonal. In contrast, the response distributions for the tablet techniques are more scattered along the diagonal. The off-diagonal errors were mostly front-back errors and they were more common for speech, than for the noise stimuli. The increased front-back confusions for speech stimuli have also been reported in other studies, both with real and virtual stimuli. Responses using the hybrid technique were less variable compared to the tablet response technique, but head-pointing localization technique outperformed the two other techniques.

At the outset, we were interested in assessing whether the order in which listeners experienced the head-pointing or tablet response technique influenced localization accuracy in the experimental conditions. The data describing the influence of order are depicted in Figure 5 where average localization errors are depicted as a function of the two types of stimuli used. The left panel depicts data for listeners who responded to Speech or Noise stimuli using the tablet response first followed by the head-pointing, whereas the right panel depicts errors for listeners who responded to the two types of stimuli using head-pointing first followed by tablet responses. A three-way analysis of variance (ANOVA), with one between-subject factor (order of response technique) and 2 within-subject factors (type of stimuli and response technique) showed a significant main effect for response techniques used ($F(2,18)=81.6$, $p<0.001$). No other main effects or interactions were significant. Thus, it appears that exposing listeners to what should be a more direct response technique first did not afford them any advantage over a group who experienced a more indirect response method first. It is perhaps surprising that the hybrid response technique resulted in lower localization errors compared to the tablet response, because in both cases, the response was the same. Somehow, turning and pointing to a location in space allowed listeners to make the necessary spatial transformation and associate a visual object with a sound with more accuracy. However, it was still not as accurate as the head-pointing technique. Listeners were required to reorient to the boresight speaker be-

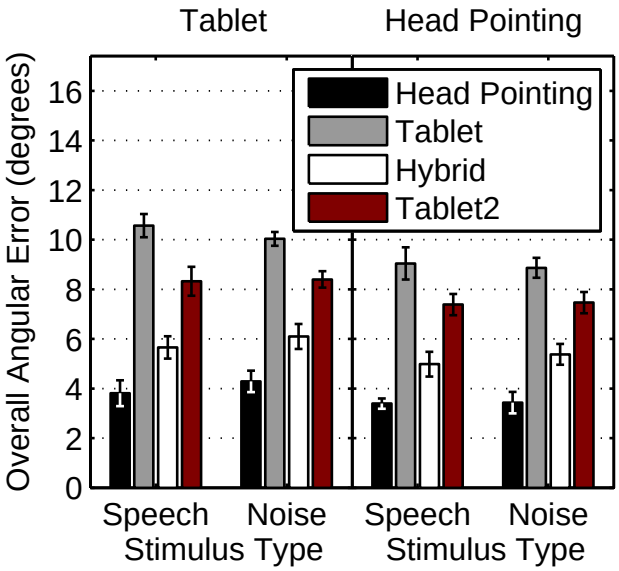

Figure 5: Average localization error plotted as a function of the two stimuli types, speech and noise. The parameters in the figures represent the three different response techniques used, as in Fig. 3. The left panel is for listeners who ran the tablet response condition first, and the right panel is for those who ran the head pointing condition first.

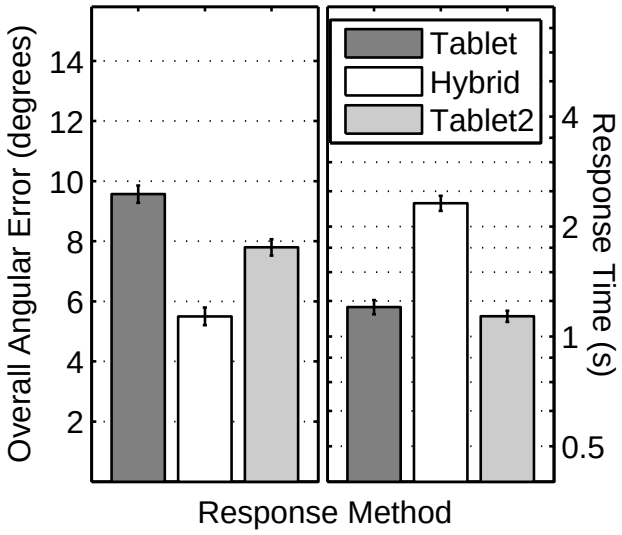

Figure 6: Average localization error (left panel) and mean response times (right panel) plotted for three response techniques, collapsed across the two stimuli type (Speech and Noise). The tablet and hybrid conditions in the figure are replotted from Fig. reffig:angular (dark grey and white bars respectively); the light grey bar depicts average error and mean response times for ten listeners who were rerun in the tablet response condition (denoted as Tablet2 in the figure). Error bars represent standard error for within-subject measures.

fore responding, which could have resulted in less accurate performance. The hybrid condition was also run as the last condition; it is possible that listeners had gained some familiarity with the tablet response technique so that better performance in the hybrid condition was merely a reflection of habituation with a response technique. In order to assess whether listeners improved in the hybrid condition due to repeated exposure/practice with the tablet response, ten of the eleven listeners from the study were re-run in the tablet response condition for a second time (the eleventh listener was unavailable). The mean accuracy and response times in this condition is plotted in Figure 6, along with data from the first tablet response condition and the hybrid condition. As shown in the figure, accuracy in the second tablet response condition reduced by approximately 2 degrees, and was significantly better than the first tablet response condition ($F(2,18)=81.6$, $p<0.001$), but still differed from the hybrid condition. However, listeners did not differ in response times between the two tablet response conditions. Thus, repeated exposure to tablet response technique facilitated learning, but the improvement did not explain all of the performance advantage observed with the hybrid response technique.

The data from the current study suggest performance degrades in listening conditions where operators are required to make a mental transformation from a sound emanating in space to an associated aerial view of the objects creating that sound (tablet response technique). The poor performance is reflected in decreased localization accuracy for the tablet response condition. When the auditory source and its corresponding visual object are co-located, listeners can easily identify the source by turning their head towards the auditory source and the visual object simultaneously (head-pointing technique). Allowing listeners to use a hybrid response technique, during which they can first turn and look at the loudspeaker and then select the appropriate one on the tablet seems to improve their localization accuracy, albeit never to the same accuracy as the head-pointing response technique condition, and at the cost of increased response times. These results suggest that associating an auditory stimuli with a visual object that do not coexist may not afford any benefits in visual search tasks. It is possible that listeners can learn to associate sound to visual objects if these sounds were perceptually different (i.e., each sound paired with a specific visual object), or if they are trained in the task.

In this task, we measured localization accuracy using three response techniques. We postulate that three possible sources can limit the performance in the task: target locations (all possible locations from which the a target sound can arise), human sensory limitation (listeners ability to localize sounds in space), response technique. From our data, it is clear that head pointing technique resulted in the best accuracy. We argue that for the target locations tested (32 loudspeakers distributed in 360 degrees azimuth), the data obtained with the head pointing technique reflects the lower bound of performance limitations; i.e., the best performance in the task given the perceptual limitations. It is possible that there might be some response transformations that listeners are required to do even in this task; nevertheless, such a transformation is well-practiced, and agrees with accuracy estimates obtained in our lab using a larger density of loudspeaker locations ([1]). Using a tablet response limited performance mainly due to response technique limitations, suggesting that GUIs used to elicit

localization responses should be used cautiously. From an auditory displays perspective, if an auditory stimulus is used to cue a location in space that is aerially displayed, our data suggests that 32 locations might be too dense of a visual field to cue. It is possible that two of the sources of performance limitations can trade-off, so that limitations due to response techniques can be offset by reducing the target location density. Therefore, GUI responses might be perfectly adequate for a more limited distribution of source locations. Further research is needed to validate our claim. Our experiment also suggested that localization accuracy was better when tablet responses were reintroduced for a second time, suggesting that exposure to a response technique could improve performance. Additional research is needed to assess whether systematic training can improve response accuracy in localization tasks using more indirect response methods.

4. CONCLUSIONS

An accumulating body of evidence has show that auditory cues play a crucial role in everyday life. Spatial auditory information can provide invaluable information to an operator, particularly when the visual channel is saturated [16]. Response times to visual targets associated with localized auditory cues have been shown to decrease [2] relative to those without auditory cues; however, in all these studies, the responses were intuitive and natural i.e., turning to look in the direction from which the sound emanated. In some Air Force operations, the transformation of auditory space to a visual object might not be straightforward, and it might require some mental transformations. The current study attempted to assess the variability associated with response techniques in a simple localization task. The findings suggest that localization errors almost double when listeners have to indicate the location from where a sound emanated using a tablet response technique, compared to a more natural head-pointing localization response. Some of the deleterious effect of making a transformation could be eliminated by allowing listeners to orient their head towards the speaker and then responding using a tablet (the hybrid response technique). Exposure to more natural response techniques did not allow listeners to perform better in conditions requiring some mental transformations. The results of the study suggests that designing auditory displays for military operations where there is not a simple 1:1 matching of the spatial locations of visual and auditory stimuli might not be particularly useful and might require additional training.

5. REFERENCES

[1] G. D. Romigh, D. S. Brungart, and B. D. Simpson, "Free-field localization performance with a head-tracked virtual auditory display," *IEEE J. Selected Topics in Sign. Proc.*, vol. 9, no. 5, pp. 943–954, 2015.

[2] D. R. Perrott, T. Sadralodabai, K. Saberi, and T. Z. Strybel, "Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target," *Human Factors*, vol. 33, no. 4, pp. 389–400, 1991.

[3] R. S. Bolia, W. R. D'Angelo, and R. L. McKinley, "Aurally aided visual search in three-dimensional space," *Human factors*, vol. 41, no. 4, pp. 664–669, 1999.

[4] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2188–2200, 1990.

[5] S. R. Oldfield and S. P. A. Parker, "Acuity of sound localisation: A topography of auditory space. I. Normal hearing conditions," *Perception*, vol. 13, no. 5, pp. 581–600, 1984.

[6] R. H. Gilkey, M. D. Good, M. A. Ericson, J. Brinkman, and J. M. Stewart, "A pointing technique for rapidly collecting localization responses in auditory research," *Behav. Res. Meth. Instr. Comp.*, vol. 27, no. 1, pp. 1–11, 1995.

[7] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, 1992.

[8] D. S. Brungart, W. M. Rabinowitz, and N. I. Durlach, "Evaluation of response methods for the localization of nearby objects," *Percept. & Psychophys.*, vol. 62, no. 1, pp. 48–65, 2000.

[9] A. W. Bronkhorst, "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.*, vol. 98, no. 5, p. 2542, 1995.

[10] R. B. King and S. R. Oldfield, "The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays," *Human Factors*, vol. 39, no. 2, pp. 287–295, 1997.

[11] S. Carlile, S. Delaney, and A. Corderoy, "The localisation of spectrally restricted sounds by human listeners," *Hearing Res.*, vol. 128, no. 1-2, pp. 175–189, 1999.

[12] A. van Schaik, C. Jin, and S. Carlile, "Human localisation of band-pass filtered noise," *Int. J. Neural Syst.*, vol. 9, no. 5, pp. 441–446, 1999.

[13] J. P. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58, no. 9, pp. 955–991, 1948.

[14] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," http://www.praat.org, 2005.

[15] J. Brindza, J. Szweda, and A. Striegel, "Wiilab," http://netscale.cse.nd.edu/twiki/bin/view/Edu/WiiMote, 2013.

[16] D. R. Begault and E. M. Wenzel, "Headphone localization of speech," *Human Factors*, vol. 35, no. 2, pp. 361–376, 1993.