EVALUATING SPATIAL-AUDITORY SYMBOLOGY FOR IMPROVED PERFORMANCE IN LOW-FIDELITY SPATIAL AUDIO DISPLAYS

Griffin D. Romigh
U.S. Air Force Research Labs
Wright-Patterson AFB, OH

Jose Villa
U.S. Army Natick Soldier RD&E Center
Natick, MA

Jason Ayers
Ball Aerospace
Dayton, OH

For decades, spatial auditory displays have been considered to be a promising technology to help fight pilot disorientation and loss of SA. Inherently heads-up, these displays can provide time-critical spatial information to pilots about navigational targets, air and runway traffic, wingman location, and even the attitude of one's aircraft without placing additional demands on the already over-tasked visual system. Unfortunately, currently-fielded auditory displays often suffer from poor spatial fidelity, particularly in elevation, due to their use of a one-size-fits-all (i.e., non-personalized) head-related transfer function (HRTF), the set of filters responsible for creating the spatial impression. The current study investigated the utility of combining a spatial cue (non-personalized HRTF) with one of two auditory symbologies, one providing both object and location information, and the other only location information. In one case, ecologically-valid sounds were paired with a particular class of visual object, and spatial cues indicated a plausible target elevation (e.g., a squeak indicated the target was a rat on the floor). In the other condition, the cue was a broadband sound, the repetition rate of which indicated target elevation (i.e., the cue provided only location information, not object information). Results indicate that target acquisition times were lower when meaningful (i.e., ecologically-valid) cues were added to non-personalized spatial cues when compared to the case in which the source-based cues provided no information about the target source. These results indicate that careful construction of auditory symbology could improve performance of cockpit-based spatial auditory displays when personalized, high-fidelity spatial processing is not practical.

Background

Because of its natural function as the body's "early-warning system," the auditory system provides an intuitive channel for portraying time-critical information. Many auditory displays leverage a listener's natural ability to rapidly identify different sound sources, and use source identification (ID) as way to alert a user to not only when, but also what type of event has occurred (e.g., different alerts for low altitude vs. traffic warnings).

Several experiments have also shown the benefits of spatial audio cues provide in visual search tasks, specifically, a reduction in visual search times compared to visual-only search conditions (Bolia et al., 1999, Perrot et al., 1996). In general, these studies have also shown that auditory-aided visual search is largely unaffected by the number of visual distractors, leading to large performance benefits for more complex visual scenes.

Displays that aim to take advantage of this spatial cueing are referred to as Virtual Audio Displays (VADs) or sometimes referred to as Spatial or 3D-Audio Displays. These displays rely on the creation of a perceptual illusion that headphone-based sounds actually originate from real-world locations in 3D space. If properly designed, VADs can have application to aircraft threat avoidance, station keeping, and navigation (Simpson et al., 2005), as well as, the more traditional use as a tool for radio speech intelligibility improvement. Despite their promise, VADs have not yet made a large impact in the aviation market, due mostly to the difficulty of achieving robust, high-fidelity spatial audio imagery on a commercial scale.

The signal processing that underlies a VAD is done by filtering a single-channel, non-spatial sound source with a pair of head-related transfer functions (HRTFs). That filtering operation results in left- and right-ear signals, which when presented over headphones can result in the perceptual illusion that the sound source was presented from a physical location out in space. Unfortunately, HRTF filters are both position- and listener- specific, meaning high-fidelity virtual auditory space can only be achieved by making electro-acoustic measurements on each listener from a large number of spatial directions. This means that commercial VAD systems, which often need to have one-size-fits-all convenience, typically have poorer fidelity than a personalized system. While lack of personalization is the major drawback of most commercial VAD technology, other compromises have also been made to save on processing power and/or battery life in some resource-constrained, real-world systems.

When non-personalized or low-fidelity HRTFs are used in a VAD, typical problems include: the perception that sources originate from inside your head (a.k.a., lack of externalization), a compression of perceived sound source elevation, and an increase in the rate of front-back reversals (the perception that sources presented in the front came from the back and vice versa) (Wenzel et al., 1993). In most applications, these perceptual shortcomings result in a decrease in the effectiveness of the VAD to accurately support its intended purpose.

The current study was designed to investigate whether the robustness of a listener's sound source identification ability could be leveraged to improve performance in an auditory-aided visual search task, when the fidelity of the spatial rendering was low.

Methods

In order to investigate whether source-ID cueing could provide a benefit for VADs with low-fidelity spatial rendering, an auditory-aided visual search task was conducted in a virtual environment with varying levels of spatial rendering quality and two auditory display symbologies that provided ID-based spatial cues.

Experimental Conditions

Ten paid listeners with normal hearing and vision participated in 24 experimental blocks over the course of three weeks. Each block consisted of 120 auditory-aided visual search trials with a fixed audio cueing condition. The audio cueing condition for each block was selected randomly from a 2 by 3 by 4 condition matrix composed of cue duration (250ms single burst, continuous), audio source type (Noise, Ecological, Click Train) and spatial rendering type (Enhanced, KEMAR, Panning, Diotic).

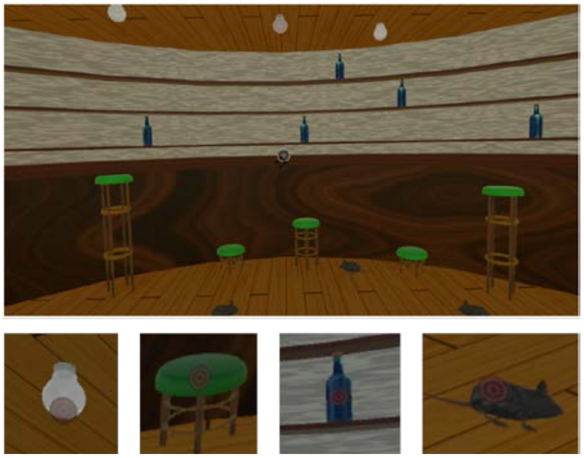

Figure 1. Virtual saloon scene used in the A/V search task (top) along with example targets from each object class (bottom).

The broadband noise stimuli were bandpass filtered between 200 Hz and 16 kHz and were independent, but statistically identical, for all target object types. This type of cue therefore provides no source-ID-based spatial information, in contrast to the ecological and click train cue types that follow. The ecological stimuli were constructed to resemble the type of auditory event a listener might expect from each of the four visual objects; electrical sparking of a light, rattling of a bottle, clanking of a barstool, squeaking of a rat. In general, all of the ecological stimuli contained spectro-temporal features sufficient to provide localization accuracy on par with the broadband noise stimuli. Conversely, random-phase click train stimuli were constructed to allow audio identification of target object classes without having any ecological validity, meaning subjects would have to learn the association between each click-train type and visual object class. The click trains were constructed by modifying the random phase click rate (100, 141, 200, and 283 Hz) and a $\sin^2$ temporal modulation window (2, 4, 6, and 8 Hz) for each class of object (rats, stools, bottles, lights, respectively). In general, these random phase click trains contain sufficient information to allow good localization; however, due to an implementation error, the click-train stimuli were lowpass filtered at 8 kHz, meaning some of the important cues for sound source localization above 8 kHz were not available.

To generate spatial audio cues, an HRTF specific to the current spatial rendering condition was loaded into slab3D and a pre-generated .wav file of the appropriate source type and duration was played through the engine. The KEMAR condition utilized a conventional non-individualized HRTF, recorded on the KEMAR mannequin as described in Romigh et al. (2015). The Enhanced condition utilized the same KEMAR HRTF after being pre-processed to exaggerate spectral cues as described in Brungart & Romigh (2009). The Panning HRTF was constructed to provide stereo panning between the left and right headphone signals based on the sound source's head-relative lateral angle. This setup means only the inter-aural level difference (ILD) cues that were relevant to sound source lateralization were present, without any high frequency monaural spectral cues, which are critical for elevation. The Diotic HRTF was constructed so that the original source signal was passed directly to both ears without any processing. This manipulation provided no spatial information, since in this condition all sound sources should appear as though they originate from the center of the listener's head.

Task Environment

Listeners were seated on a rotating stool in the Spatial Hearing Anechoic Research Chamber (SHARC) at Wright Patterson AFB, OH. An audio-visual virtual environment was presented via an HTC Vive VR headset and a pair of Sennheiser HD280 headphones. The Vive allows 6-DOF motion and also includes a tracked wand to enable cursor-based pointing within the scene. Spatial audio rendering was accomplished using slab3D (Miller & Wenzel, 2002)an open-source audio rendering engine that allows incorporation of custom HRTFs and has been shown to produce virtual sound sources that permit localization accuracy on par with free-field sources (Romigh et al., 2015).

The virtual environment was created in Unity3D, a game engine for developing interactive 3D virtual environments. The environment resembled a 360^o saloon scene (top panel of Figure 1) and consisted of a cylindrical room partitioned into four distinct regions in elevation (i.e., floor, bar, wall, ceiling). In each elevation region, 30 instances of a single class of object were scattered randomly throughout the region at all azimuths; light objects occupied the ceiling region from +36 to +18 degrees in elevation, bottle objects occupied the wall region from +18 to 0 degrees in elevation, stool objects occupied the bar region from 0 to -18 degrees in elevation, and rat objects occupied the floor region from -18 to -36 degrees in elevation.

Experimental Task

Each trial started when the listener pulled the trigger to “shoot” the large bullseye in the front of the visual scene by aiming a wand-slaved crosshair cursor. Then, a visual target was presented in the form of a semi-transparent bullseye placed randomly in front of one of the 120 scene objects. The transparency of the target bullseye was manipulated to subjectively equalize the salience of all target objects and make it less likely that a visual target could be identified in the visual periphery. Simultaneously, a virtual audio cue was presented from the location of the visual target, and the task of the subject was to locate and shoot (aiming a wand-slaved crosshair) the visual target as quickly and as accurately as possible. The first shot aimed within 10 degrees of the visual target scored as a hit and the timing of the shot was recorded as the response time.

Results and Discussion

Average response times for all conditions are shown in Figure 2. Results for short duration, “Burst” stimuli and continuous stimuli and shown in the left and right panels, respectively. In general, response times for the burst and continuous stimuli were similar. The biggest differences appear to be for Noise stimuli and/or the Panning rendering condition. This suggests that when some cues for sound source elevation are available (i.e. in the Enhanced and KEMAR rendering conditions and/or with Ecological or Click Train stimuli) the additional information provided by dynamic head-motion cues and a longer observation window do not reduce search times.

With the Noise stimuli, response times increased with decreasing spatial rendering quality, as expected, rising from 2 seconds in the Enhanced condition to over 6.5 seconds in the Diotic condition. In contrast, the Ecological stimuli were less affected by rendering condition, increasing from just under 2 seconds in the Enhanced condition to roughly 4 seconds in Diotic condition. The Click Train stimuli fell in between the Noise and Ecological conditions, which could have resulted from both decrease in localizability caused by its reduced bandwidth, or because the mapping of the click train parameters to elevation (or source type) was less intuitive that the Ecological stimuli. The fact that a difference is seen between the Ecological and Click Train stimuli in the Diotic condition suggests the latter.

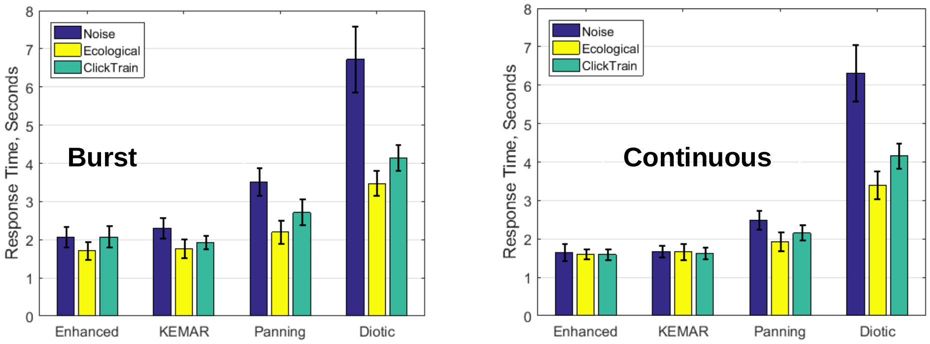

Figure 2. Average response times for each experimental condition. Error bars represent their 95% confidence intervals.

Comparing the results across the conditions, it appears that providing source-ID based elevation cues can provide increasing benefit in terms of reduced search times as the fidelity of the spatial rendering cues goes down. The largest benefit is therefore found when no rendering-based elevation cues are available (e.g. in the Panning and Diotic conditions); however, since the performance benefit between Noise and Ecological stimuli goes up from the Panning to the Diotic condition, it suggests another non-spatial cue is being used (e.g. a benefit from a reduced valid set-size).

Figure 3 shows head-tracking elevation data for all Burst trials. Each panel represents a different experimental condition, as indicated, and colors are used to identify the target object type and target elevation range (Purple-Lights, Cyan-Bottles, Yellow-Stools, Red-Rats). Dramatic differences are apparent for the Ecological and Noise stimuli. Even in the Enhanced rendering condition where average response times are fairly similar, the ecological stimuli clearly resulted in more definitive head movements, as can be seen by the clear separation of tracks with different target object types (i.e., tracks with different colors). This suggests very different search strategies are employed when different sources of spatial information are available.

Conclusion

The current study investigated the benefit of providing source-ID based spatial cues in addition to traditional spatial rendering cues in an auditory-aided visual search task. Response time results indicate that the benefit of adding source-ID cues goes up with decreasing fidelity of the spatial rendering, and may not be influenced by stimulus duration and/or presence dynamic head-motion cues. Head tracking results indicate that different search strategies are employed when source-ID based cues are available.

Acknowledgements

This work was partially supported through funding from the Air Force Office of Scientific Research (Grant #16RHCOR366).

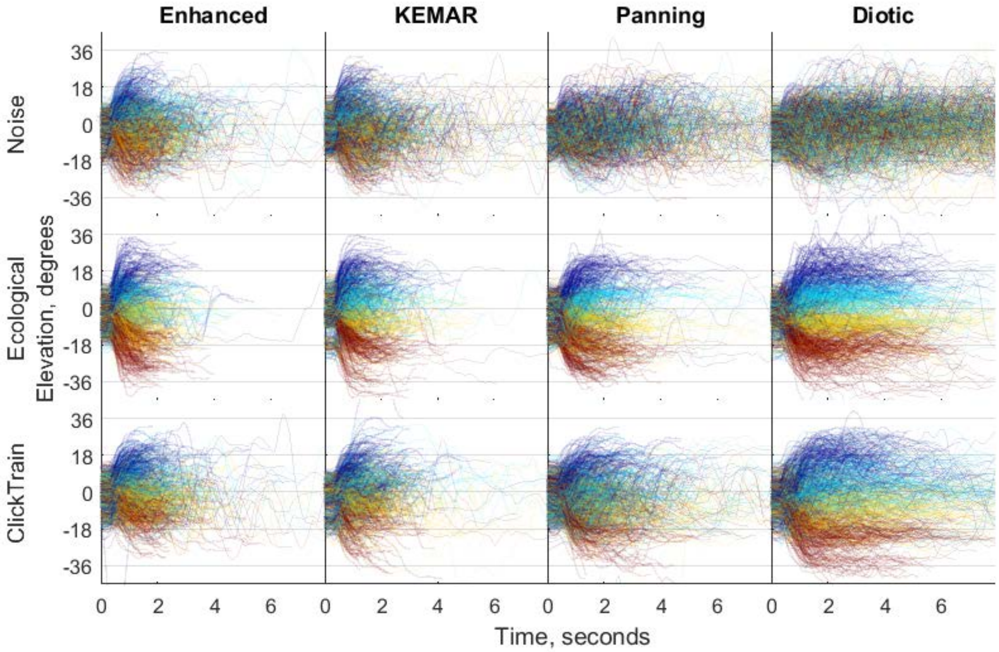

Figure 3. Headtracking data for all trials from the short duration "Burst" trials. Tracks show the elevation component of the head-orientation as a function of time. Each panel shows a single experimental condition. Colors indicate the target object type (Purple – lights, Cyan – Bottles, Yellow – Stools, Red – Rats).

References

Bolia, W. S., D'Angelo, W. R., & McKinley, R. L. (1999). Aurally aided visual search in three-dimensional space. Human Factors, 41, 664-669.

Brungart, D. S. and Romigh, G. D.(2009) "Spectral HRTF enhancement for improved vertical-polar auditory localization," *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 305-308. doi: 10.1109/ASPAA.2009.5346479

Miller, J. D., Wenzel, E. M. (2002) Recent Developments in SLAB: A Software-Based System for Interactive Spatial Sound Synthesis, Proceedings of the International Conference on Auditory Display, ICAD 2002, Kyoto, Japan, pp. 403-408

Perrott, D. R., Cisneros, J., McKinley, R. L., & D'Angelo, W. R. (1996). Aurally aided visual search under virtual and free-field listening conditions. Human Factors, 38, 702–715.

Romigh, G. D., Brungart, D. S., Simpson, B. D. (2015). Free-field localization performance with a head-tracked virtual auditory display. IEEE Journal of Selected Topics in Signal Processing, 9, 943-954

Simpson, B.D.; Brungart, D.S.; Gilkey, R.H.; McKinley, R.L. (2005) Spatial Audio Displays for Improving Safety and Enhancing Situation Awareness in General Aviation Environments. In New Directions for Improving Audio Effectiveness, pp. 26-1 – 26-16.

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