

## **SIMPLE VISUALIZATIONS ENHANCE SPEAKER IDENTIFICATION WHEN LISTENING TO SPATIALIZED VOICES**

Ryan Kilgore and Mark Chignell  
Dept. of Mechanical & Industrial Engineering  
University of Toronto  
Toronto, ON, Canada

Spatial audio has been demonstrated to enhance performance in a variety of listening tasks. The utility of visually reinforcing spatialized audio with depictions of voice locations in collaborative applications, however, has been questioned. In this experiment, we compared the accuracy, response time, confidence in task performance, and subjective mental workload of 18 participants in a voice-identification task under three different display conditions: 1) traditional mono audio; 2) spatial audio; 3) spatial audio with a visual representation of voice locations. Each format was investigated using four and eight unique stimuli voices. Results showed greater voice-identification accuracy for the spatial-plus-visual format than for the spatial-and mono-only formats, and that visualization benefits increased with voice number. Spatialization was also found to increase confidence in task performance. Response time and mental workload remained unchanged across display conditions. These results indicate visualizations may benefit users of large, unfamiliar audio spaces.

### **INTRODUCTION**

The use of spatial audio has been shown to enhance auditory perception (Ericson & McKinley, 1997; Abouchacra, 2001; Bolia, 2001) and to facilitate visual searches (Perrott, 1996; Bronkhorst, 1996; Bolia, 1999). Spatialization of talkers' voices has also been demonstrated to aid in the cognition of audio conference events and to enhance listeners' later recall of who said what (Baldis, 2001; Kilgore et. al., 2003). Investigations regarding spatial communication, however, have shown little additional performance benefit from visually depicting apparent voice locations within spatialized audio environments via graphical user interfaces, or GUIs (Baldis, 2001; MacDonald, 2002). In fact, some researchers have presented design guidelines explicitly advising practitioners against developing such visualizations, stating "GUIs are not well suited to audio spaces," because "audio communication does not demand visual attention" (Singer, et. al., 1999). It is important to note, however, that these studies have involved co-located workers (likely to be very familiar with each other's voices) or small numbers of potential voices ( $N = 4$ ), and thus may be insensitive to possible benefits of visualizations in this context.

Based on our own experiences in developing and field-testing spatialized audio communication tools—both with and without GUIs—we hypothesize that visualizations may in fact present useful aids for voice identification, particularly in environments with large numbers of unfamiliar talkers. The goal of this study was to investigate whether ability to identify voices could be enhanced through a combination of spatialization and visual representation of voice locations, with the intent of informing future display design.

### **EXPERIMENTAL METHODS**

The purpose of this experiment was to assess the benefits of depicting voice locations in a spatialized, multi-talker voice-identification task. Dependent variables included the accuracy and response times of listeners' voice identifications and measurements of participants' confidence in task performance. Mental workload was also assessed, using a modified version of the NASA Task Load Index (Hart & Staveland, 1988).

#### **Hypotheses**

Our expectation was that the use of spatialized voices, coupled with visual representations of the audio space, would result in a greater ability to identify voices in comparison to traditional mono audio, or spatialized audio that was not supported with visual feedback. We also hypothesized that the benefits of combined spatial audio and visualization would increase with increasing number of voices. Thus, Hypothesis 1 was that performance would be more accurate (without being any slower) with visual feedback, and Hypothesis 2 was that the advantage of visual feedback would be greater for environments with eight voices than with four voices (leading to a significant interaction of number of voices and format on accuracy).

#### **Participants**

A mix of eighteen students between the ages of 18 and 28 (mean age = 22.5) participated in this experiment. All participants reported normal hearing in both ears and were paid CAD\$15 for their time.

**Design**

The study featured a mixed ANOVA design, with two independent variables: the number of voices presented and the format of the audio/visual display.

Two levels of the Number of Voices variable were manipulated within-subjects: eight voices and four voices. Participants received each level twice—for a total of four experimental blocks—with order counterbalanced using an ABBA design. Both the eight-voice and the four-voice conditions contained an even mix of male and female voices.

Three Format conditions were examined using a between-subjects unrelated design: Mono, Spatial, and Spatial+Visual. In the *Mono* format, voices were unspatialized and accompanied by an alphabetized list of voice names (Figure 1, below). In the *Spatial* format, voices were lateralized to appear at evenly spaced locations along the listener's interaural axis—using Interaural Time Difference (ITD) cues—and were accompanied by an alphabetized list of voice names. In the *Spatial+Visual* format, lateralized stimuli identical to those of the *Spatial* condition were used. However, the *Spatial+Visual* format also included a depiction of the relative locations of stimuli voices (Figure 2, below).

**Apparatus**

A single laptop computer was used to present stimuli and record subjects' responses. Stimuli for the Mono format consisted of 44,000Hz, 16-bit sound files obtained from the Coordinate Response Measure (CRM) corpus, developed by Bolia et. al., (2000). This corpus consists of 2,048 unique phrases of the form "Ready (CALL SIGN), go (COLOR) (NUMBER) now," with four male and four female voices each reading 256 Call Sign, Color, Number combinations (8×4×8).

Auditory stimuli used in the *Spatial* and *Spatial+Visual* conditions also originated from the CRM corpus. For these stimuli, Matlab was used to introduce Interaural Time Differences (ITDs), or phase shifts, between the stereo audio channels to simulate non-individualized 'within-the-head' lateralizations of the stimuli. All auditory stimuli were normalized for volume (RMS) and presented to participants using a Sennheiser HD-500 headset.



Figure 1. Interface for the Mono and Spatial formats. The vertical arrangement of voices is alphabetized and provides no spatially-encoded information.

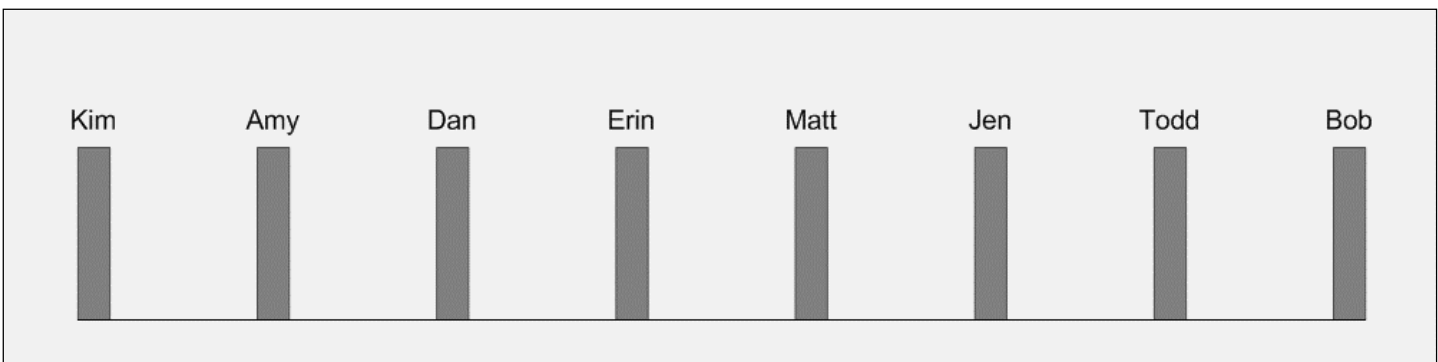


Figure 2. Interface for the Spatial+Visual format. The left-to-right arrangement of voices in the display corresponds to the relative apparent locations of the speakers' voices.

**Procedure**

Before experimentation, participants practiced using the response interface. Participants also learned to recognize the eight voices used in the study through approximately 10-15 minutes of scripted, self-paced training that included performance feedback from the administrator. All participants were required to achieve at least 75% voice-identification accuracy before proceeding.

During the experiment, participants received four treatment blocks, each consisting of 40 trials and lasting approximately eight minutes. Participants were instructed to respond to target stimuli by selecting the Color/Number pair of the stimuli and then indicating the name of the person they believed to be speaking the phrase. Following each block, participants completed a questionnaire consisting of a confidence scale and the six NASA-TLX subscales.

**RESULTS**

**Correct Voice Identification**

Significant main effects were found for both Number of Voices,  $F(1,15) = 70.55, p < .001$ , and Format,  $F(2,15) = 9.54, p = 0.02$ . As seen in Figure 3, participants correctly identified more stimuli for the four-voice condition than for the eight-voice condition. Post hoc analyses using Scheffe's method were used to assess the source of the significant difference ( $p < .05$ ) between the three Format conditions. As predicted by Hypothesis 1, voice identification performance for the Spatial+Visual format was found to differ significantly from both the Mono and Spatial formats (mean differences of +28.3% and +21.1%, respectively). No significant difference was found between the percentage of correctly identified voices for the Mono and Spatial formats.

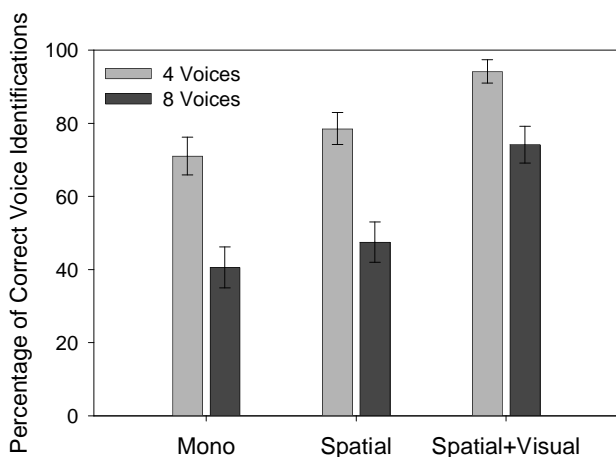


Figure 3. Percentage of correctly identified voices, as a function of Number of Voices and presentation Format. Error bars indicate one standard error of the mean in each direction.

With respect to Hypothesis 2, a borderline significant Number of Voices by Format interaction effect was found,  $F(2,15) = 3.57, p = 0.05$ . The performance improvement effect for the Spatial+Visual format tended to be greater for the eight-voice condition than for the four-voice condition (Figure 3). This result suggests that, as hypothesized, the visual cues provided by the Spatial+Visual interface offer greater utility with an increasing number of talkers.

**Response Time**

Analysis of natural log transforms of response times showed significant main effects for Number of Voices,  $F(1,15) = 122.16, p < .001$ , but not Format ( $F < 1$ ). Additionally, a significant interaction effect (Number of Voices by Format),  $F(2,15) = 6.73, p < 0.01$ , was found. As can be seen in Figure 4, mean response times were shorter overall for the four-voice conditions. Also, there was a consistent effect of presentation format on response time, but only for the four-voice condition.

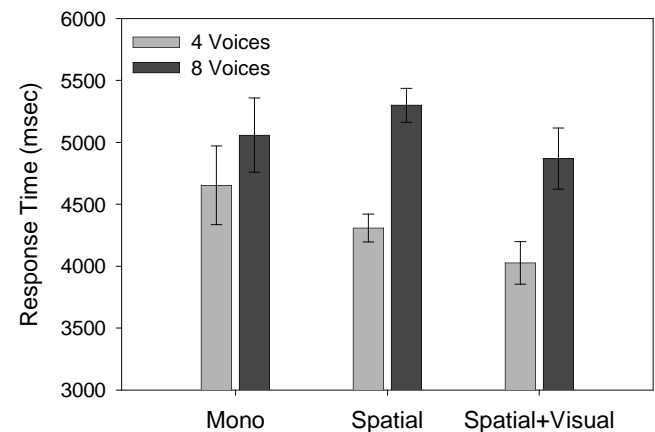


Figure 4. Mean response time for target stimuli, as a function of Number of Voices and presentation Format. Error bars indicate one standard error of the mean in each direction.

**Confidence**

With respect to confidence scores, significant main effects were found for the Number of Voices,  $F(1,15) = 97.17, p < .001$ , and for Format,  $F(2,15) = 3.61, p = 0.05$ . These two effects are depicted in Figure 5, on the following page. Confidence was greater for the four-voice condition than for the eight-voice condition across all three audio/visual presentation formats. Also, for both Number of Voices levels, confidence ratings were greater for the Spatial and Spatial+Visual formats than the Mono format. No significant Number of Voices by Format interaction effect was found ( $F \approx 1$ ).

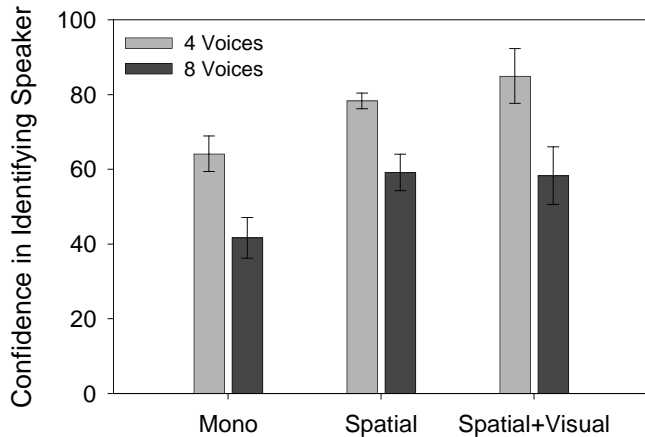


Figure 5. Mean confidence scores for the Speaker Identification task, as a function of Number of Voices and presentation Format. Error bars indicate one standard error of the mean in each direction.

**Mental Workload**

Overall workload scores (unweighted sums of the six subscale scores) showed a significant main effect for Number of Voices,  $F(1,15) = 66.09, p < .001$ , but not for Format. Nor was there a Number of Voices by Format interaction ( $F < 1$  in both cases). As can be seen in Figure 6, overall workload was greater for eight voices than for four voices across all three Formats. Overall workload ratings tended to be greater for the Mono format than for both the Spatial and Spatial+Visual formats, but post hoc analyses using Scheffe's method did not show these differences to be statistically significant ( $p > .05$ ). Application of ANOVA to the six workload subscale responses showed similar results, with a main effect of Number of Voices ( $p < .05$ ) in all cases but no effect for Format nor for the Number of Voices by Format interaction ( $F \approx 1$  in all cases).

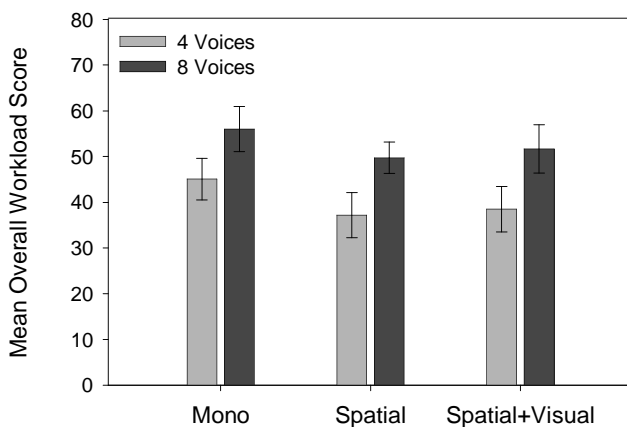


Figure 6. Mean Overall (summed) Workload scores, as a function of Number of Voices and presentation Format. Error bars indicate one standard error of the mean in each direction.

**DISCUSSION**

This research examined the benefits of adding simple visualizations to a spatialized, multi-talker audio space. Participants' accuracy in identifying voices was found to increase significantly with the inclusion of visual representations of voice location, particularly when there were a larger number of voices present (eight vs. four). Confidence in voice identification performance, on the other hand, was found to increase with spatialized audio, regardless of whether or not visualization was included. Measures of mental workload and response time were not found to vary significantly between formats, although both increased with increasing voice number.

These findings suggest that simple visual representations of voice locations aid listeners in disambiguating voices, building upon a number of performance benefits previously identified for spatial audio. Given these results, we suggest that spatialized voice streams, coupled with simple visual depictions of voice locations, may be particularly useful in supporting large collaborative groups, especially when participants may have only limited familiarity with each others' voices.

**REFERENCES**

Abouchacra, K. (2001). Binaural Helmet: Improving speech recognition in noise with spatialized sound. *Human Factors*, 43 (4), 584.

Baldis, J. (2001). Effects of spatial audio on memory, comprehension, and preference during desktop conferences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vol. 3, 2001, 166-173.

Bolia, R. (1999). Aurally aided visual search in three-dimensional space. *Human Factors*, 41 (4), 664.

Bolia, R. S., W. T. Nelson, M. A. Ericson, and Simpson, B. D. (2000). A speech corpus for multitalker communication research. *J. Acoust. Soc. Am.*, 107 (2), 1065-66.

Bolia, R. (2001). Asymmetric performance in the cocktail party effect: implications for the design of Spatial Audio Displays. *Human Factors*, 43 (2), 208.

Bronkhorst, A. (1996). Application of a three-dimensional auditory display in a flight task. *Human Factors*, 38 (1), 23.

Ericson, M.A., and McKinley, R. L. (1997). The intelligibility of multiple talkers separated spatially in noise. in Robert H. and Timothy R. Anderson (Eds.) *Binaural and Spatial Hearing in Real and Virtual Environments*, Gilkey, NJ, Lawrence Erlbaum Associates, 701-724.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati (Eds.), *Human Mental Workload*. North Holland: Elsevier Science Publishers, 139-183.

Kilgore, R. M., M. Chignell, and Smith, P.W. (2003). Spatialized audioconferencing: what are the benefits? *Proceedings of the 2003 Centre for Advanced Studies Conference on Collaborative Research*, 111-120.

Perrott, D. (1996). Aurally aided visual search under virtual and free-field listening conditions. *Human Factors*, 38(4), 702.

MacDonald, J. (2002). Intelligibility of speech in a virtual 3-D environment. *Human Factors*, 44(2), 272.

Singer, A., D. Hindus, L. Stifelman, and White S. (1999). Tangible progress: less is more in Somewire audio spaces. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 104-111.