

# Listening to Unfamiliar Voices in Spatial Audio: Does Visualization of Spatial Position Enhance Voice Identification?

*Ryan Kilgore*

*Mark Chignell*

*Interactive Media Lab – University of Toronto, Toronto, Canada*

[r.kilgore@utoronto.ca](mailto:r.kilgore@utoronto.ca)

[chignell@mie.utoronto.ca](mailto:chignell@mie.utoronto.ca)

## **Abstract**

The use of spatial audio to present voices from unique locations around a listener's head has been demonstrated to enhance the perception and cognition of auditory events for a variety of listening tasks. This paper describes experimental research to determine whether a combination of spatialization and simple visual representation of voice locations further builds upon the known benefits of spatial audio by aiding listeners in learning to recognize completely unfamiliar voices. A combined spatial audio and visual display format resulted in significantly stronger learning of unfamiliar voices than either mono audio or spatial audio displays with no visual depictions of voice locations. This benefit, however, was only present when the audio space contained a large number of voices. Insight from this research will be useful in informing the design of future multi-modal interfaces for collaborative environments and in developing methodologies for experimentally evaluating potential interfaces for such applications.

**Key words:** Spatial audio, visualization, voice identification, audioconferencing

## **1. Introduction**

This paper describes a psychoacoustic study that investigated the performance effects of a user interface that combined spatialized audio with visual depictions of voice locations in a completely unfamiliar auditory environment. Insights derived from this experiment will be used to inform the design of interfaces for collaborative audio environments that better support large, unfamiliar groups of talkers, as well as to develop future methodologies for experimentally evaluating potential interfaces for collaborative audio applications.

## **2. Background**

### **2.1 Spatial Audio and Remote Collaboration**

The use of spatial voice streams has been shown to yield a number of benefits in the perception of information transmitted via the auditory modality. Prominent among these benefits is the increased intelligibility of speech signals in noisy surroundings (Ericson and McKinley, 1997), or when multiple voices are speaking simultaneously (Drullman and Bronkhorst, 2000; Abouchacra, 2001; Bolia, 2001). These benefits are the result of spatial audio's facilitation of auditory scene analysis, the process by which the brain reduces a sample of simultaneous audio streams into individual constituent sounds (Bregman, 1990).

The distinct location of voices also aids in cognitive processing of conference events, in part by facilitating the task of speaker identification. A study of spatial voice streams in an audioconference-style listening task (Baldis, 2001) demonstrated that the locational separation of conferee voices increased listeners' ability to later recall which voice said what. The use of spatialized conferee voices was also shown to result in significant subjective benefits, namely an increase in the perceived overall comprehension of conference events and a reduction in perceived attention requirements for speaker identification. Spatial presentation of voices was also found to be significantly preferred by listeners to a traditional, mono (non-spatial) presentation format (Baldis, 2001). Finally, it has been shown that providing subjects with the ability to control the location of conference participants' voices decreases subjective measurements of the difficulty and attentional effort associated with identifying voices and significantly reduces the number of confusions made between similar voices (Kilgore et. al., 2003).

## **2.2 Visual Depiction of the Audio Space**

Previous research regarding communication-based listening tasks has shown little to no benefit in providing listeners with visual depictions of spatial voice locations (Baldis, 2001; MacDonald, 2002). These findings suggest that such visual representations aren't necessary in collaborative audio spaces. Some researchers have presented design guidelines that explicitly point practitioners away from including visualizations, stating "GUIs are not well suited to audio spaces," because "audio communication does not demand visual attention" (Singer, et. al., 1999). However, it is important to note that these studies have involved co-located workers (likely to be familiar with each other's voices) or small numbers of voices within the audio space (N = 4).

A previously reported study (Kilgore & Chignell, 2005), examined the benefits of visually representing an audio space when listeners were identifying voices that they had already been trained to recognize (to a 75% recognition accuracy criterion). The study's methodology was based on the real-world scenario where people would have met prior to a conference and thus have a reasonably good, but imperfect, idea of who was speaking. The study compared the impact of three audio/visual interfaces on listener's ability to identify voices: a standard mono audio presentation, similar to traditional teleconferencing tools; a spatial audio presentation, with voices presented from different apparent locations around the listener's head; and a combined spatial and visual presentation, where spatial audio was coupled with a visual representation of voice locations. Participants' accuracy in identifying voice stimuli was found to be roughly the same for mono and spatial formats, but significantly increased with the inclusion of a visual representation of voice locations. This benefit was found to be greater when the audio space was comprised of eight voices than when it was comprised of four.

The results of this previous study suggest that low-cost, non-video representations of voice locations support the disambiguating of voices when listeners already had some familiarity. What was not investigated, however, was the extent to which a combined spatial and visual representation of voice locations aided listeners in learning new, completely unfamiliar voices within an audio environment. This is the focus of the current study presented in this report.

## **3. Psychoacoustic Study**

The purpose of this current psychoacoustic study was to assess the performance effects of visual depictions of voice locations in the identification of completely unfamiliar voices. Dependent variables included both the overall accuracy and response times of listeners' voice

identifications, as well as changes in accuracy over sequential blocks of experimental stimuli as voice familiarity increased. Subjective measurements of participants' confidence in their task performance were obtained via a questionnaire, and cognitive demand (mental workload) was also assessed using a modified version of the NASA Task Load Index (Hart & Staveland, 1988).

### **3.1 Hypotheses**

Our expectation prior to experimentation was that the use of spatialized voices, coupled with a visual representation of the listener's audio space, would result in a greater ability to identify and learn voices (as compared with traditional mono audio, or spatialized audio that was not supported with visual feedback). We also hypothesized that the benefits of combining spatial audio with visual representation would be more significant for a larger number of voices. With this experiment, we hoped to confirm that combining spatialized audio with a visual interface would provide benefits in completely unfamiliar voice spaces similar to those already demonstrated in partially familiar voice spaces.

### **3.2 Design and Participants**

We implemented a between-subjects design, with the number of voices that were presented and the format of the audio/visual display manipulated as independent variables. Twenty-seven participants listened and responded to a series of short auditory stimuli. The participants were a mix of undergraduate and graduate students between the ages of 18 and 33 (mean age = 23), recruited using email lists for the department of Mechanical and Industrial Engineering at the University of Toronto. All participants reported normal hearing in both ears and were able to distinguish between sounds localized to the left and right side of the head in a brief hearing test administered prior to the experiment. The experiment lasted approximately one hour and participants were paid (Canadian) \$15 for their time.

Two levels of the Number of Voices independent variable were investigated: eight voices and four voices. Both the eight-voice and four-voice conditions consisted of an equal number of male and female voices (four and two of each gender, respectively). Participants completing the four voice condition performed a set of two experimental treatments, with a unique set of voices used in each treatment. As the corpus of auditory stimuli used in this experiment contained only eight unique voices, participants completing the eight-voice condition performed only one experimental treatment.

Three different Format conditions were investigated. In the *Mono* format, voices were presented equally to both ears (similar to wearing a headset in a traditional audio conference). The Mono format also included a visual component that listed stimuli voice names in alphabetical order. For the *Spatial* format, voices were lateralized to appear at distinct, evenly spaced locations along the listener's interaural axis, using Interaural Time Difference (ITD) cues. Voices remained in a constant location for each participant, but the relative ordering of the voices was randomized between participants. The Spatial format also included the same visual component used for the Mono format—an alphabetical list of voice names. In the *Spatial+Visual* format, auditory stimuli used were identical to those of the Spatial condition. However, the visual component of the Spatial+Visual condition corresponded to the relative locations of the stimuli voices (Figure 1, on the following page).

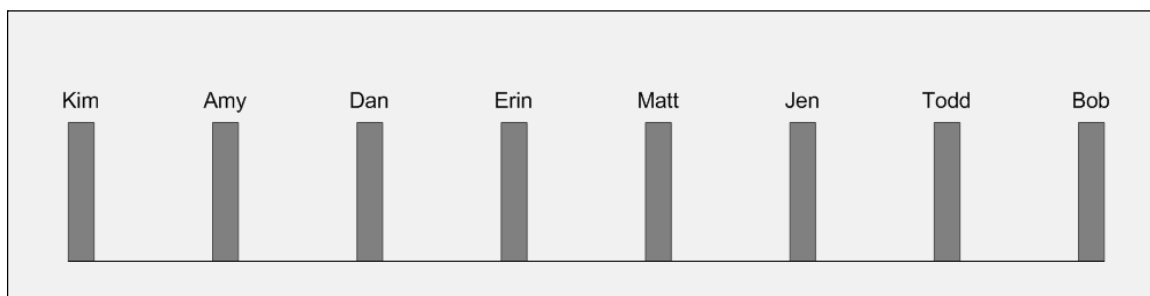


Figure 1. Interface for the Spatial+Visual condition. The left-to-right arrangement of voices in the display corresponds to the relative apparent locations of the eight speakers' voices.

### 3.3 Apparatus and Procedure

The experiment was conducted in a quiet room using a single laptop to both present stimuli and to record and time subject responses, which were input via mouse selection of radio buttons. In contrast to an earlier study (Kilgore and Chignell, 2005), there was no training on voice identification prior to the experimental sessions, and the experimental interface in this study provided explicit performance feedback to participants. Following each response, a short audio clip was presented to the participants, indicating whether their response was correct or incorrect. Additionally, a visual indication of both the subject's response and the correct response was displayed for a brief duration after each response was entered.

Auditory stimuli used in this experiment consisted of 44,000Hz, 16-bit sound files obtained from the Coordinate Response Measure (CRM) corpus, developed by Bolia et. al. (2000). The CRM corpus consists of 2,048 unique phrases of the form "Ready (CALL SIGN), go to (COLOR) (NUMBER) now," for example "Ready Baron, go to Green Three now." Eight unique voices—four male and four female—read each of 256 possible Call Sign, Color, Number combinations ( $8 \times 4 \times 8$ ), and each file is roughly 1.5 seconds in duration. Stimuli for the Mono format were obtained directly from the CRM corpus. Stimuli for the Spatial and Spatial+Visual formats were created from copies of these audio files, with Matlab used to introduce Interaural Time Differences (ITDs), or phase shifts, between the left and right audio channels to simulate various 'within-the-head' lateralizations of the sound source.

Participants were presented with either one or two treatments—depending on whether they were selected for the four-voice or eight-voice condition as discussed previously—each consisting of four sequential blocks of 40 stimuli each, for a total of 160 stimuli trials. Participants were instructed to respond immediately following each stimulus by selecting the Color/Number pair that was uttered and then indicating the name of the voice they believed to be speaking. Use of the Color/Number pair of the CRM stimuli provides a secondary task requiring participants to attend not only to identifying the voice that is speaking but also to the informational content of the message they are hearing—a quality similar to real-world interactions. After each experimental treatment, participants completed a questionnaire form consisting of a subjective confidence scale and the six NASA-TLX subscales.

## 4. Results

*Overall Correct Voice Identification:* Natural log transforms of participant performance over the course of the entire experimental treatment (overall percentages of correctly identified target voices for the 160 treatment stimuli) were examined for each condition using a between

subjects ANOVA with two factors (display format and number of voices). The results of this analysis are illustrated in Figure 2.

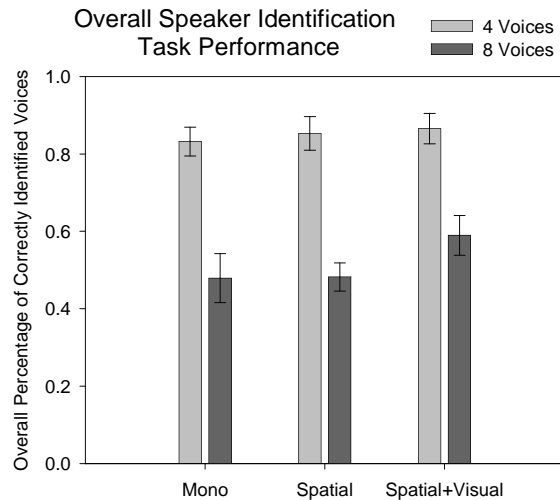


Figure 2. Percentage of correctly identified voice stimuli, as a function of Number of Voices and presentation Format. Error bars indicate one standard error of the mean in each direction

As expected, a significant main effect was found for the Number of Voices ( $F[1,30] = 64.52$ ,  $p < .001$ ) on participants' correct identification of stimuli voices. As seen in Figure 2, participants correctly identified more stimuli for the four-voice condition than for the eight-voice condition, across all three audio/visual formats. The effect of Format on *overall* speaker identification performance was not statistically significant ( $F[2,30] = 1.60$ ,  $p = .22$ ), nor was the Number of Voices by Format interaction ( $F < 1$ ). However, in the eight-voice condition, the mean percentage of correct responses trends to be greater for the Spatial+Visual format (59% accuracy) than for both the Mono and Spatial formats (both 48% accuracy). Accuracy in the four-voice condition was high across all three of the formats (Mono, Spatial, and Spatial+Visual) with 83%, 85% and 86% accuracy, respectively.

*Learning of Unfamiliar Voices Over Time:* The main focus of this study was on how strength of learning varied across the three format conditions. A mixed design analysis of variance was performed using the two between-subjects factors, along with natural log transforms of participant performance data aggregated into quartiles of sequential 40-trial blocks as a within-subject factor. The results of this analysis can be seen in the two graphs shown in Figure 3, on the following page. A significant main effect on accuracy was found for the within-subjects experimental Block factor ( $F[3,30] = 61.15$ ,  $p < .001$ ), with speaker identification task performance improving as participants progressed through the experiment while receiving continual performance feedback, as can be seen in Figure 3. A significant main effect was also found for the Number of Voices being presented ( $F[1,30] = 68.21$ ,  $p < .001$ ). As Figure 3 shows, speaker identification accuracy was higher for the four-voice condition (versus the eight-voice condition) for each of the four experimental blocks and across all three format conditions. The audio/visual presentation format was not found to have a significant effect on block-by-block speaker identification performance in this analysis ( $p > .05$ ). However, while speaker identification performance for each block appeared unaffected by Format in the four-voice condition (as seen on the left side of Figure 3), performance for the Spatial+Visual condition tended to be greater than that of the Spatial and Mono conditions

when eight-voices were presented to the listeners (as seen on the right side of Figure 3), although this effect did not reach statistical significance.

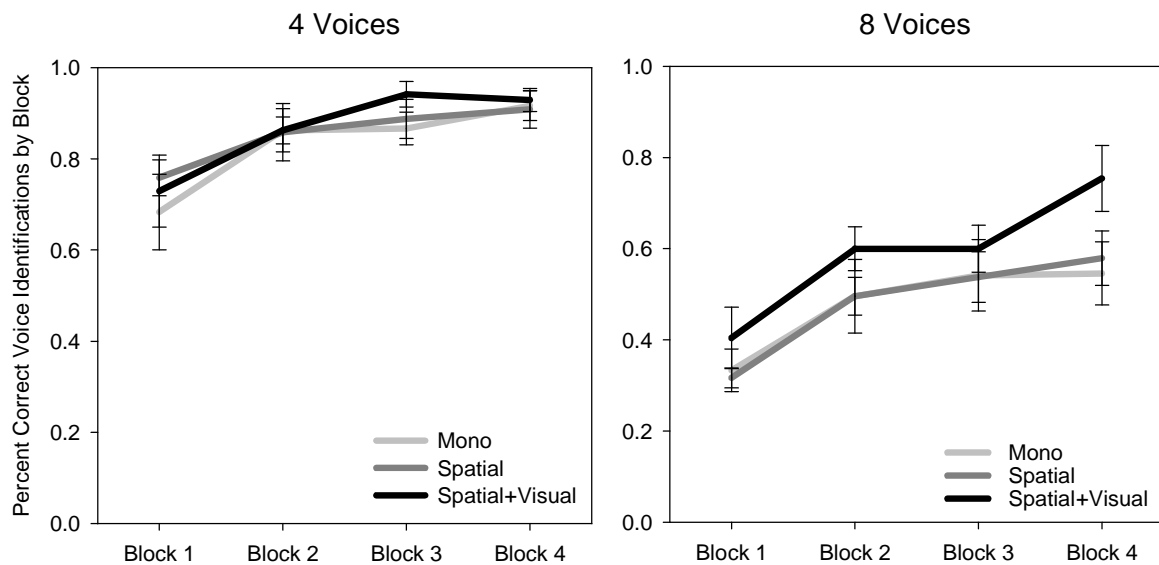


Figure 3. Percentage of correctly identified voices, aggregated into sequential 40-stimuli blocks. Error bars indicate one standard error of the mean in each direction.

Since voices were completely unfamiliar to the participants when the experimental sessions began, we were interested in how the strength of voice learning varied by condition. The strength of learning for each participant was calculated using the following procedure: First, the 160 trials for each participant were aggregated into 16 groups of 10 trials each, and the number of correct voice identifications out of the 10 possible was calculated for each of those groups. Since the learning curve was assumed to follow the usual exponential distribution (e.g., Hancock and Bayha, 1992), the natural logs of the trial group numbers were calculated and regressed against the number of accurate responses for the group in order to fit the learning curve. The correlation of the resulting fitted regression lines was then used as a measure of the goodness of fit to the learning curve and served as a measure of the consistency of learning over the course of the experimental session.

We then removed data sets for eight treatments where participants had fitted correlations of less than 0.33 (i.e., no more than 10% of the variance accounted for by the learning effect), on the grounds that they weren't consistently learning voices over the course of the experiment and thus should not be considered in the comparison of learning effects between the conditions. Two of these data sets were from the Mono format, three were from the Spatial format, and three were from the Spatial+Visual format. For the remaining 28 data sets where participants showed consistent learning, performance data was then aggregated into two major blocks: the first 120 trials versus the last 40 trials, on the grounds that learning was occurring in the earlier trials, but that the final set of trials would best represent an approximation of asymptotic performance, as suggested previously in Figure 3.

A repeated-measures analysis of variance was performed for this data set, using two experimental blocks (trials 1-120 and 121-160, respectively) as a within-subject factor. A borderline three-way interaction was found between Number of Voices, audio/visual Format, and Block ( $F[2,22] = 3.05$ ,  $p = .067$ ). We further investigated the source of this possible

interaction by running separate two-way analyses of variance for the four-voice and eight-voice conditions. These analyses, illustrated in Figure 4 below, showed that no interaction was occurring between audio/visual Format and experimental Block for the four-voice condition ( $F < 1$ ), where performance was already quite high. However, a significant Format by Block interaction was found for the eight-voice condition ( $F[2,10] = 5.43, p = .025$ ). As shown in Figure 4, speaker identification during the final block of trials for the eight-voice condition was significantly greater for the Spatial+Visual format than for the Mono and Spatial formats, with performance approaching that seen for the four-voice conditions. This finding indicates the powerful value of the Spatial+Visual interface on participants' learning of unfamiliar voices, particularly in larger audio environments.

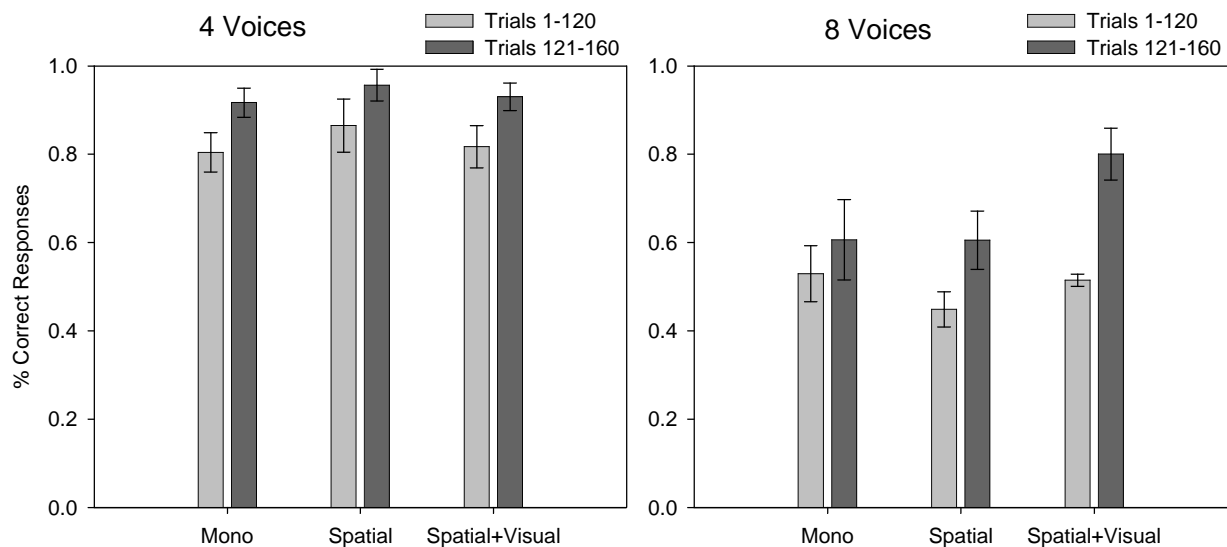


Figure 4. Speaker Identification task performance, with low-learning subjects removed. Error bars indicate one standard error of the mean in each direction.

## 5. Discussion

This study demonstrates the benefit of providing a visual interface to aid listeners in learning to identify completely unfamiliar voices when using spatialized audio. However, this benefit was found only to occur with larger numbers of voices, being present in this study for the eight-voice condition but absent in the four-voice condition. Given the results of the present study, it is likely that earlier studies (Baldis, 2001; MacDonald, 2002) did not find a benefit when including a visual representation of the audio space because they used only four stimuli voices. It seems a large number of voices are needed before visualization provides a useful supplement to audio spatialization for voice identification.

The findings of this experiment—using initially unfamiliar voices, but with performance feedback during the experimental session—are consistent with those of an earlier study (Kilgore and Chignell, 2005), where participants were trained to a 75% level of accuracy in voice identification prior to running the experimental session. Thus, not only do visual interfaces assist in the learning of unfamiliar voices, but as participants become familiar with the voices the visual interface continues to support better performance. Based on the beneficial effect of the visual interface with eight—but not four—voices, we expect that the benefit of using a visual interface will increase as the number of voices is increased above eight. Given that business audio conferences often have many more than eight participants, future research with larger voice sets is warranted. Another concern for future investigations

is the additional workload generated by using the visual interface. While there were no significant differences in workload reported between the conditions in the present study, there was a tendency for more workload to be reported when using the visual interface. The question of whether or not visualization creates a significant workload increase when larger numbers of voices are present should also be examined in future research.

In conclusion, visualization of spatial position enhances voice identification when a sufficiently large number of voices is used (with the minimum threshold being somewhere between five and eight voices, based on the present results). Given the well-known advantages of spatialized audio in audio conferencing, it is clear that future collaborative systems should incorporate both spatialized audio and visual interfaces that support the user's mental representation of the conference, particularly when there are more than a few voices, some of whom are unfamiliar.

## 6. Acknowledgements

The authors would like to thank Brandon Kilgore for his help in developing the software used in this experiment to present CRM stimuli and record subject responses.

## 7. 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.S., W.T. Nelson, M.A. Ericson, and B.D. Simpson, (2000). A speech corpus for multitalker communication research. *J. Acoust. Soc. Am.*, 107 (2), 1065-66.
- Bregman, AS., (1990), *Auditory Scene Analysis*. Cambridge: MIT Press.
- Brungart, Douglas S., and Brian D. Simpson, (2003) Optimizing the spatial configuration of a seven-talker speech display, *Proceedings of the 2003 International Conference on Auditory Display*, Boston, MA, USA, July 6-9, 188-191.
- Drullman, Rob and Adelbert W. Bronkhorst, (2000), "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *J. Acoust. Soc. Am.* 107(4), April, 2224-2235.
- 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.
- Hancock, W.M., and Bayha, F.H., (1992). The Learning Curve, in Salvendy, G. (Ed.), *Handbook of Industrial Engineering*. 2nd Edn., Wiley, New York, 1585-1598.
- 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., and Chignell, M., (2005). Simple Visualizations Enhance Speaker Identification when Listening to Spatialized Voices, *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*.
- MacDonald, J., (2002). Intelligibility of speech in a virtual 3-D environment. *Human Factors*, 44(2), 272.
- Singer, A., D. Hindus, L. Stifelman, and S. White, (1999). Tangible progress: less is more in Somewire audio spaces. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 104-111.