

The Vocal Village: Enhancing Collaboration with Spatialized Audioconferencing

Ryan Kilgore
University of Toronto
r.kilgore@utoronto.ca

Mark Chignell
University of Toronto
chignell@mie.utoronto.ca

Abstract: Traditional methods of voice communication fail to adequately address the needs of users engaged in remote collaborative activities. While some emerging technologies offer greater support for these activities, they often do so at a considerable price, making them unsuitable solutions for widespread usage. To address this demand, we have developed the Vocal Village, an enhanced audioconferencing application that uses VoIP technology to join members of distributed collaborative groups in immersive and personalizable spatialized audio environments. This paper describes the Vocal Village system and also presents the findings of a preliminary study, in which it was determined that the methods of personalizable spatialization used within the Vocal Village significantly improved participants' subjective measures of audioconference performance.

Introduction

Reliance on group communication over a distance has increased with the accelerated use of personal computing and a greater frequency of collaboration in dynamic and cross-organizational project teams (Mills, 1999). This, coupled with an increasing demand for e-learning solutions, has created a need for effective and efficient applications for remote group work. Although an ever greater amount of collaboration at a distance is being conducted, tools to support activities such as e-learning are often difficult to use and unpopular (Kraut et. al., 2002). Traditional methods of synchronous communication, such as telephones and speakerphones, do not adequately support situation awareness in large groups of people. With these audio-only technologies, individual voices can become difficult to distinguish and users are forced to maintain mental models of the collaborative space—who is present, who is coming, who just left—without external aids. Additionally, voice quality tends to be quite poor in these environments, exacerbating difficulties in determining who is speaking, especially in large and unfamiliar groups such as those encountered in many e-learning settings.

Alternatives to traditional methods of voice communications, such as videoconferencing, attempt to increase the perception of *presence* within remote collaborative environments by accommodating visual interactions between remote users. However, these systems often require peripheral hardware and large amounts of dedicated bandwidth, making them prohibitively expensive for widespread use by the general public. Instead, what is needed is a lightweight collaborative tool that increases the communication effectiveness and improves the satisfaction and experience of distributed groups in a manner that is suitable for use both in corporate settings *and* in the home. We have developed the Vocal Village—a tool that supports enhanced forms of remote collaboration through spatialized audioconferencing—in an attempt to address this need.

Spatialized Audio in Collaborative Environments

The term 'spatialized audio' refers to sounds that are perceived to originate from discrete location in space. Similar to depth perception via binocular vision, spatial hearing occurs as the result of our ability to perceive minute differences between the patterns of sound waves striking the left and right eardrum. These differences are the result of numerous factors, including the distances between a sound source and the left versus the right ear (Inter-aural Temporal Difference, or ITD) and relative volume differences due to occlusion by the head itself (Inter-aural

Intensity Difference, or IID) (Gilkey and Anderson, 1997). Additionally, the unique shape of the outer ear—or pinna—modifies the frequency spectrum of sounds entering the inner ear in a manner that is dependent upon the direction and distance of a sound event, through what is known as a Head-Related Transfer Function (HRTF) (Blauert, 1997).

The use of spatialized voice streams within collaborative audio environments has been shown to yield numerous psychophysical benefits, including the increased intelligibility of speech signals in noisy surroundings (Ericson and McKinley, 1997). This benefit is significant, given that online audio environments—including those employed by many existing e-learning technologies—not only suffer from transmission-related noise, but are also prone to noise caused by end-user equipment, such as low-quality hardware or improperly used microphones (Watson and Sasse, 2000). Additionally, when multiple conferees speak simultaneously in traditional, monaural audioconferences, their voices emanate from a single location in space. This causes the voice streams to mask each other, making the separation and comprehension of individual voices a difficult task at best. However, speech intelligibility in multi-talker listening tasks can be increased by spatializing the voice streams (Drullman and Bronkhorst, 2000). The discrete location of sound events facilitates the process of auditory scene analysis by aiding the brain in reducing a sample of simultaneous audio streams into individual constituent sounds (Bregman, 1990). In effect, this spatialization allows each listener to selectively attend to a single voice and ignore others, a useful feature in noisy or crowded collaborative environments.

The distinct location of voices in a spatial auditory environment has also been demonstrated to aid in the cognition of collaborative events, in part by facilitating the task of speaker identification. A study of high-fidelity spatialized voice streams in an audioconference-style listening task (Baldis, 2001) demonstrated that the locational separation of conferee voices increased listeners' ability to later identify the speaker of transcribed conference statements. The use of spatialized conferee voices also resulted 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. Spatialized audio was significantly preferred by listeners to a traditional, monaural presentation format. These benefits of spatialized audio are of particular relevance to e-learning applications, as they directly impact the extent to which virtual collaborative environments are able to successfully imitate face-to-face learning experiences.

The Vocal Village

Our aim is to further enhance the quality of collaboration and knowledge dissemination between dispersed individuals by creating light-weight audioconferencing tools that more effectively reproduce the benefits of traditional forms of direct, face-to-face communication. Such tools are now technologically and economically feasible due to advancements in the capabilities of Voice over Internet Protocol (VoIP) systems. To meet our goal of increasing the communication effectiveness and improving the satisfaction and experience of distributed groups, we have developed the Vocal Village system.

The Vocal Village (available for download at www.vocalvillage.net) is a real-time audioconferencing application that uses VoIP technology to connect collaborative groups over the Internet. Unlike other VoIP applications, the Vocal Village enhances traditional audioconference environments by binaurally presenting auditory location cues that cause the voices of individual participants to appear as if they are coming from different positions in space, distributed from left to right around the listeners' head. This spatialized presentation format more closely simulates the acoustical qualities of face-to-face collaboration, allowing Vocal Village users to feel as though they are sharing an actual physical auditory space. The ecological nature of this collaborative environment results in a more satisfying auditory experience—and potentially a greater feeling of presence—than is possible with traditional monaural telephone or speakerphone systems. The Vocal Village system is capable of spatially distributing multiple voices using only standard stereo sound outputs. Because of this, the system may be run on a personal computer or laptop with no additional hardware other than a standard set of stereo headphones and a microphone.

A Visual Interface for Voice Collaboration

Within the Vocal Village environment, users are able to independently arrange incoming voice streams across the horizontal plain and view a graphical depiction of their collaborative audio space. This graphical user interface (Fig. 1) displays contextual information to Vocal Village users, showing who is present in the audio space and where they are located, as well as providing a visual cue when participants enter and leave the collaborative space. Additionally,

the spatial location of individual voices, coupled with this graphical depiction, serves as a cue to disambiguate speaker identity. This can be very useful in ad hoc or unfamiliar collaborative settings where people have not had the opportunity to learn each others' voices. If, for example, a user hears a voice that they do not recognize coming from the left side of their head, they can refer to the visual depiction of the conference to aid in determining who is speaking.

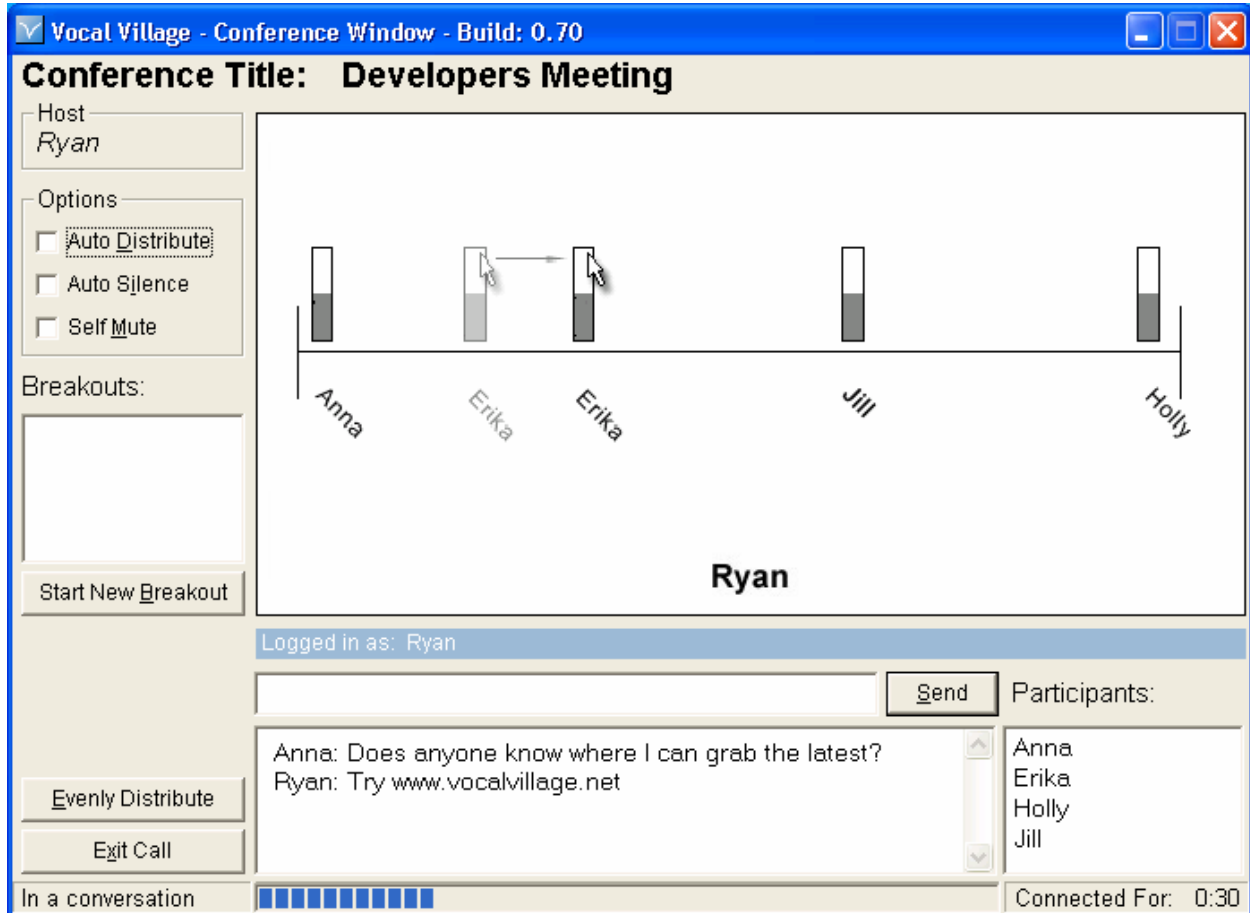


Figure 1: The Vocal Village graphical user interface

The visual depiction also serves as a means for controlling the virtual auditory environment, allowing listeners to change the apparent location and volume of other participants' voices in a natural click-and-drag style of interaction. This scheme provides individual listeners with the ability to "personalize" their audio space by controlling where the voice each person they are talking to appears to be located and how loud that voice is. The visual feedback of the Vocal Village interface enhances the usability of this novel auditory space by facilitating in the maintenance of users' situation awareness, providing listeners with information that may otherwise be missed or ambiguous in traditional audio-only environments.

In addition, the visual interface of the Vocal Village also provides users with the ability to communicate with each other through an integrated text messaging system. This supplementary mode of communication accommodates private text messages between two Vocal Village users, as well as messages that are publicly presented to the entire collaborative group.

Enhanced Collaborative Functionality

In addition to accommodating rich, spatialized voice communications, the Vocal Village provides other enhancements to traditional audio collaboration that are of particular relevance to e-learning applications: breakouts and annotated recordings of conversations.

Vocal Village breakouts are sub-conferences within a main conference, where collaborators can go to participate in semi-private discussions. When located in a breakout, a participant's voice is only heard by other members of that breakout. However, members of a breakout are able to simultaneously monitor the main conversation at a "dimmed" volume level. This allows users participating in a breakout to periodically monitor or selectively attend to the events of main conference. Within the Vocal Village, participants are able to traverse efficiently between breakouts and the main conference, allowing for ad hoc collaboration to occur between small groups participating in larger collaborative environments. In an e-learning setting, such breakouts could facilitate group-work on projects or assignments. For example: small groups of students could collaborate in multiple independent breakouts, while simultaneously seeking and receiving guidance from an instructor who remains in the main conference. In such a scenario, the instructor is free to move between individual breakouts, monitoring groups' progress and offering advice, similar to walking from table to table in a real-world classroom. If desired, the instructor can also spontaneously return to the main conference and lecture to all of the breakout groups simultaneously. These breakouts are intended to accommodate small-group interactions in a naturally dynamic and evolving manner that is simply not possible with traditional forms of audio communication.

An integrated tool for annotating and reviewing conversations will also be included as a feature in the forthcoming version of the Vocal Village software. This system will be particularly useful within e-learning applications because it will serve as an efficient tool for bridging the gap between synchronous and asynchronous forms of collaboration, providing lasting archives of Vocal Village meetings. These recordings will consist of temporally linked audio and textual archives of conversations, integrating the spoken word and text messaging of Vocal Village collaborations. Additionally, we have developed a "player" application for listening to and viewing digitally recorded conferences. This player allows participants to perform quick visual scans of textual messages and use them to jump to key points in the recorded conversation for audio playback. This system uses the built-in Vocal Village text messaging system as a means of annotating and searching audio recordings. In an e-learning context, a student participating in a lecture within the Vocal Village would be able to use this feature to leave text messages as personal bookmarks or notes, making later review of the lecture audio a more efficient process. For example, if a physics instructor were to discuss a new formula that a student wished to later review, that student could type a message such as "*new formula*" into the indexing system while listening to the live discussion. The message is then stored as a time-stamped event in synch with the recorded audio conversation. Later, using the playback system, the student would be able to scroll through a time-ordered list of the text message annotations that they made during the lecture. Clicking on the message "*new formula*" would begin playback of the audio loop from the point in time when the annotation was made, eliminating the need to re-listen through the entire lecture and streamlining the note-taking process.

Technical Specifications

The Vocal Village system was specifically designed to offer a more rich and satisfying experience than traditional telephone and speakerphone systems without requiring the elaborate and expensive hardware common to many existing systems that support enhanced forms of collaboration at a distance, such as videoconferencing. Because of this, the Vocal Village is a suitable environment for general and cross domain e-learning uses, ranging from corporate enterprise to consumer applications.

The system uses a client/server architecture that allows audioconference participants to access the Vocal Village through individual clients running on their personal computers. Users' voices travel as compressed mono sound streams to the central Vocal Village server, which then combines all of the conferee voices into spatialized signals that are in turn sent back to the individual clients as a single compressed stereo sound stream. This client/server architecture used by the Vocal Village minimizes the bandwidth requirements placed on individual users, as each client receives only a single stereo sound stream from the server, regardless of the number of conference participants. The relatively low bandwidth demand of Vocal Village clients allows users to participate in large audioconferences from home via broadband cable modem or DSL Internet connections.

Spatialization of voices is achieved within the Vocal Village by manipulating mono voice streams to incorporate Interaural Time Differences (ITDs) and Interaural Intensity Differences (IIDs), the two major binaural cues for lateralizing sounds in space (Bernstein, 1997). This simplified form of “within-the-head” spatialization—referred to as *lateralization*—does not allow for highly accurate modeling of sound locations. For example, it is impossible to model elevation or depth using only these two cues. However, it is a sufficient method for synthesizing distinct voice locations in the horizontal plane. Use of only ITD and IID cues saves on computational processing requirements and eliminates the need for individual calibration required by more sophisticated spatialization systems due to differences in the shapes of people’s heads and outer ears. This emphasis on limiting process requirements supports a key goal of the Vocal Village system, which is to accommodate nearly instantaneous spatialized communication across a variety of platforms. As such, the minimal processing demands placed on Vocal Village clients may be satisfactorily met by many existing wireless and handheld devices, allowing future implementations of the Vocal Village to accommodate mobile participants.

Preliminary Findings

A preliminary study was performed to investigate the cognitive benefits associated with use of the Vocal Village in a simulated audioconference listening task (Kilgore, et. al., 2003). The purpose of this preliminary study was to determine if the low-fidelity, “within-the-head” spatialization techniques used in the Vocal Village were sufficient to yield performance benefits comparable to those exhibited for high-fidelity spatialized audio by (Baldis, 2001). Also investigated was whether providing users with the ability to choose the apparent location of conference participants within the virtual auditory space (personalization) was capable of further enhancing any such performance benefits low-fidelity systems might achieve.

Results from this research showed that, similar to high-fidelity spatialization, the use of low-fidelity, lateralized voices significantly improved participants’ subjective measures of conference performance. These measures included an increase in users’ confidence in their ability to identify speakers and in their perceived comprehension of conference events. Additionally, the perception of both the difficulty and the attentional requirements associated with identifying speakers during the conference were found to be significantly lesser for the Vocal Village spatialized audio format than for the traditional monaural format. Providing subjects with the ability to control the apparent location of conference participants (personalization) resulted in the greatest benefit to both of these measures. Interestingly, when given the chance to arrange the voices in space, 18 out of the 22 experimental participants purposefully separated two particular conferees and explained that they did this because their voices sounded similar. Confusions between these two conferees were significantly reduced when the audio was presented in this personalized format. However, unlike high-fidelity spatialization, the implementation of low-fidelity, “within the head” lateralization in this study did not significantly improve, nor impair, objective measurements of conference performance, including listeners’ ability to remember who said what or to recall the general viewpoints of individual conferees (Kilgore, et. al., 2003).

Future Work

The existing version of the Vocal Village spatialized audioconferencing environment is undergoing a process of redesign, in an attempt to more effectively address the needs of a variety of collaborative uses and settings. Although we continue to refine the system as a fully standalone communication package, we are also examining ways of integrating the spatialized audio and enhanced collaborative functionality of the Vocal Village into other existing collaborative applications, such as archival telecasting or electronic whiteboard systems. Currently, our effort is also being directed towards the development of a Vocal Village client that is specifically targeted to address the unique demands of e-learning environments. We hope to work closely with members of the e-learning community to design a tool that better facilitates a high quality of collaborative effectiveness and greater user satisfaction for distributed individuals and project teams engaged in educational activities.

Acknowledgements

This research has been supported by Bell University Laboratories and IBM's Centre for Advanced Studies, Toronto Lab. We would like to thank Paul Smith from the IBM Centre for Advanced Studies for all his help and input throughout the duration of this project. We would also like to thank the entire Vocal Village Team for their invaluable efforts.

References

- Baldis, Jessica, (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, pp. 166-173.
- Bernstein, Leslie R., (1997) "Detection and discrimination of interaural disparities: Modern earphone-based studies" in Binaural and Spatial Hearing in Real and Virtual Environments, Gilkey, Robert H. and Timothy R. Anderson Eds., New Jersey: Lawrence Erlbaum Associates, pp. 117-138
- Blauert, Jens, (1997) Spatial hearing: the psychophysics of human sound localization, Rev. ed., MIT Press, Cambridge, MA.
- Bregman, A. S., (1990) Auditory Scene Analysis. Cambridge: MIT Press.
- 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), pp. 2224-2235.
- Ericson, M.A., and R. L. McKinley, (1997) "The intelligibility of multiple talkers separated spatially in noise," in Binaural and Spatial Hearing in Real and Virtual Environments, Gilkey, Robert H. and Timothy R. Anderson Eds., NJ, Lawrence Erlbaum Associates, pp. 701-724.
- Gilkey, Robert H. and Timothy R. Anderson Eds., (1997) Binaural and Spatial Hearing in Real and Virtual Environments, New Jersey: Lawrence Erlbaum Associates.
- Kilgore, Ryan M., Mark Chignell and Paul W. Smith, (2003) "Spatialized Audioconferencing: what are the benefits?" Proceedings of the 2003 conference of the Centre for Advanced Studies conference on Collaborative research, 2003, pp. 111-120 .
- Kraut, R. E., Fussell, S. R., Brennan, S. E., & Siegel, J. (2002). Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. In P. Hinds & S. Kiesler (Eds.) Distributed Work, Cambridge, MA: MIT Press, pp. 137-162.
- Mills, Kevin L., (1999) "Introduction to the electronic symposium on computer-supported cooperative work." ACM Computing Surveys, Vol. 31, No. 2, pp. 105-115.
- Watson, Anna and M. Angela Sasse, (2000) "The good, the bad, and the muffled: the impact of different degradations on Internet speech," Proceedings of the eighth ACM international conference on Multimedia, pp. 269-276.