# Exploring Spatialized Audio & Video for Distributed Conversations

**Kori Inkpen, Rajesh Hegde, Mary Czerwinski, Zhengyou Zhang**
Microsoft Research, 1 Microsoft Way
Redmond, WA, USA
{kori | rajeshh | marycz | zhang}@microsoft.com

**ABSTRACT**
Previous work has demonstrated the benefits of spatial audio conferencing over monophonic when listening to a group conversation. In this paper we examined three-way distributed conversations while varying the presence of spatial video and audio. Our results demonstrate significant benefits to adding spatialized video to an audio conference. Specifically, users perceived that the conversations were of higher quality, they were more engaged, and they were better able to keep track of the conversation. In contrast, no significant benefits were found when mono audio was replaced by spatialized audio. The results of this work are important in that they provide strong evidence for continued exploration of spatialized video, and also suggest that the benefits of spatialized audio may have less of an impact when video is also spatialized.

**Author Keywords**
Spatial audio, spatial video, video conferencing, audio conferencing, distributed meetings, collaboration.

**ACM Classification Keywords**
H5.3. [**Group and Organization Interfaces**]: *collaborative computing, CSCW, evaluation/methodology*.

**General Terms**
Design, Experimentation, Human Factors, Measurement

**INTRODUCTION**
Recently, the demand for better quality teleconferencing with remote participants has risen rapidly, especially with efforts to reduce travel costs, leverage the global workforce, and increase productivity. Effective communications and collaboration fundamentally require multimodal interaction and rich media, including both verbal and non-verbal communications and data sharing.

Today's audio conferencing (whether traditional telephony or Voice-over-IP) is essentially monaural and is most suitable for one-to-one communications. Monaural applications are not as applicable to other scenarios, such as multi-party conferences, one-to-many meetings and many-to-many meetings. One major problem in these scenarios is that a participant at one end has difficulties in identifying who is talking at the other end and comprehending what is being discussed. One reason is that the voices of multiple participants are intermixed into a single audio stream.

Adding a video stream to a monaural audio stream is one way to overcome the limitations of monaural audio conferencing. Traditionally, desktop video conferencing systems show the current speaker, thereby reducing the cognitive load in determining who is talking and also conveys visual cues of the current speaker. However, as we know from face-to-face meetings, meeting dynamics also depend on reactions of other participants to the speaker. Previous work has shown that providing one video stream of the current speaker and another showing a panoramic view of all participants was highly valued [2].

In this paper, we explore the impact of spatial audio and video on users' experience in a multi-way videoconference. Spatialized audio is the use of sound effects to create the illusion of sound sources placed in 3D space. Similarly, spatial video involves the placement of video streams in 3D space. We developed a system which utilizes spatialized audio and video. Unlike most multi-party video conferencing systems where all remote parties see the same video of the local participant, our system uses a dedicated camera, display, microphone and speaker for each remote participant, providing participants with the correct awareness of who is looking at whom and who is speaking to whom. Our monitor/camera/speaker/mic sets are similar to the Hydra system [6] in that they serve as proxies for the remote participants; however, the Hydra units were much smaller in size. Because we use a 22 inch wide screen display, distributed participants have a higher fidelity experience, allowing them to better perceive others' visual cues (facial expressions, body language, etc).

Our work shows significant benefits from adding spatial video to an audio conference, however, users did not perceive strong benefits from spatialized audio. This result suggests that despite benefits of spatial audio reported in previous work [1, 5, 7], spatial video has a strong impact on users' perceived engagement, their ability to keep track of the conversation, and perceived quality of the conversation.


*CSCW 2010,* February 6–10, 2010, Savannah, Georgia, USA.

### Benefits of Video and Spatial Audio

The benefits of video to support distributed collaboration have been mixed in previous research. For example, early work by Ochsman and Chapanis [3] revealed no significant benefit of using video as compared to an audio only link. However, in 1993, based on a case study of video conferencing, Isaacs and Tang [4] articulated several benefits of video, including the ability to show understanding; forecast responses; and use gestures to emphasize a point. Sellen [6] also found similar benefits in her study of several different conferencing systems.

The benefits of spatial audio for audio conferencing have also been explored by several researchers. Baldis [1] explored non-spatial audio in comparison to two high-fidelity, spatial audio conditions. Her results demonstrated significant performance benefits for the two spatial audio conditions. She also examined the impact of including static images to represent the people in the conference. While there were no significant benefits in terms of performance, most of the participants indicated that they preferred having the visual representations.

Kilgore, Chignell and Smith [5] also explored spatial audio conferencing, but instead used simulated sound spatialization using standard stereo sound outputs. Contrary to Baldis's results, they found no significant increase in memory; however, participants did prefer the spatial audio over monaural, and participants also perceived memory and voice identification benefits.

More recent work by Yankelovich et al., [7] found that high-fidelity stereo audio improves task performance and speaker differentiation, and increases a sense of social presence. Audio quality also had a huge impact. Users' subjective ratings of effort, quality, and feeling of presence were also enhanced with spatial audio.

Our work differs from previous research in three main ways. First, previous research on spatialized audio has primarily focused on users *listening* to multi-party conversations [1, 5, 7]. Although one of Kilgore et al.,'s [5] conditions involved a real conversation between users, it was only a two-way audio conference. In contrast, we wanted to gather users' perceptions when they were *actively participating* in a multi-way, distributed conversation.

A second key difference is that much of the previous work on spatial audio has focused solely on audio [5, 7], and has not examined the impact of adding video to the spatial audio. Given that one of the benefits of spatialized audio is speaker identification, we wanted to explore whether live video would be as good, if not better, than spatial audio at helping participants keep track of the conversation.

A third difference is that much of the previous research on spatial audio has used high-end systems to situate the audio streams [1, 7]. Like Kilgore et al., [5] we were interested in a lower-fidelity approach to spatial audio using separate microphones which transmit signals to separate speakers.

## STUDY

The goal of our study was to explore users' reactions to the use of spatial audio and video for distributed conversations.

### Participants & Setting

Twenty-four participants (8 groups of 3) were recruited to participate in our study. Four groups were all male, one was all female, and three were mixed (total of 8 females). Participants in each group knew each other and participated in meetings together regularly. All participants were given a small gratuity for their participation.

### System

We used an in-house multi-party video conferencing system for this study. A schematic diagram of this video conferencing system is shown in Figure 1.

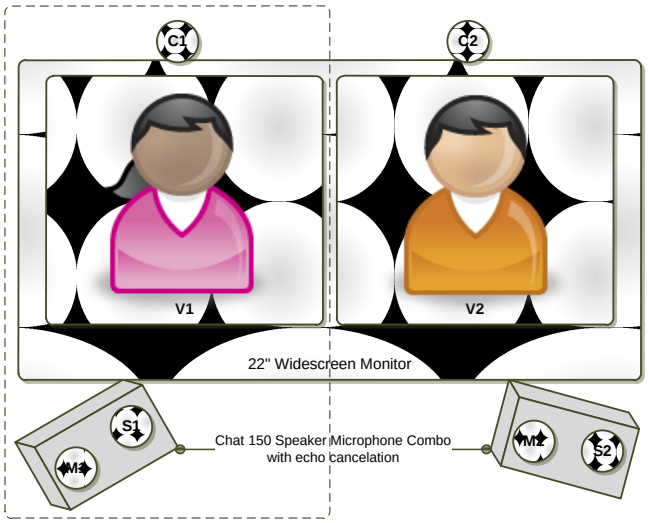


Figure 1: Conferencing software and hardware setup with cameras (C1,C2), microphones (M1,M2), speakers (S1,S2), and videos of remote participants (V1,V2).

The system consisted of a desktop computer in each room connected to a 22 inch wide display with a resolution of 1600x1200. There were two cameras mounted on the display (C1 and C2) and the video frame rate for the system was 30 fps. For audio, we used a Chat 150 device that contained a microphone and a speaker with embedded echo cancellation. When the 3-way video conference started, our software automatically organized the participants into a virtual circle, simulating three users sitting around a table. Doing this preserved the relative gaze for each participant. Participants sat approximately 2 ft. away from the display, similar to a typical desktop setup. The software created a "conferencing unit" for each remote participant consisting of a video window, camera, microphone and a speaker (V1, C1, M1 and S1 shown inside the dotted box in Figure 1). This organization ensured that any given remote participant has their remote eye (camera), ears (microphone), mouth (speaker) all on the same side of the display.

### Procedure

We employed a 2 X 2 repeated measures design with two factors: audio (mono vs. spatial) and video (with vs.

without) (see Table 1). Condition order was counterbalanced using a Latin-Square design.

**Table 1. Descriptions of the four conditions in the study.**

| Mode | Description |
|---|---|
| Mono Audio Only (MAO) | Audio only conference. No video shown. Speech was played back through one speaker positioned in the center of the display. |
| Stereo Audio Only (SAO) | Audio only conference. No video shown. Left remote participant's speech was played back on the left speaker (S1) and the right remote participant's speech was played back through the right speaker (S2) |
| Mono Audio Video (MAV) | Same as MAO, but video of both remote parties was shown this time. |
| Stereo Audio Video (SAV) | Same as SAO, but video of both remote parties was shown this time. |

Upon arrival, participants were introduced to the study and the equipment and then separated into three different rooms, one for each individual. Participants were given a series of "current event" topics and asked to discuss these topics with other members of their group for **three minutes**. The topics were always discussed in the same order to ensure an equal distribution across the experimental conditions. The decision to discuss current events was modeled after similar work by Sellen [6], where her participants were asked to participate in several informal debates. The topics discussed in our study included: nationalizing health care; illegal immigration; auto industry bailout; and state income tax.

In order to easily switch between the conditions, we created a task sequencer tool. This tool consisted of four buttons MAO, SAO, MAV, and SAV representing each of the four conferencing modes. Clicking on any of these buttons activated the conference in that mode. We did not tell the participants what each mode meant in order to minimize any bias. In order to start the conference, the researcher instructed the participants to click on a specific button.

After each condition participants were asked to complete a short questionnaire, inquiring about different aspects of the conversation such as conversation quality, level of engagement, and how easy it was to keep track of the conversation. After completing all four conditions the participants were given an additional questionnaire which asked them to describe any differences they perceived across the conditions and the impact of those differences.

## RESULTS

Although the rating data provided by the participants at the end of each condition were ordinal, we used repeated measures ANOVAs to analyze the data on quality, engagement, and tracking, in order to accommodate the factorial design of the study. Analyzing the data without taking the main factors into account using non-parametric statistics revealed similar trends to the traditional ANOVA. We also utilized qualitative data from the questionnaires to provide additional insights on the conditions.

### Audio Quality

Our participants did not perceive any significant differences in audio quality across the four conditions (Friedman: $X^2(3)=6.71$, $p=.08$). Therefore, any differences in audio would likely be a result of the spatialization.

### Quality of the Conversation

Participants felt that adding video significantly improved the quality of the conversation ($F_{1,23}=5.61$, $p=.03$, $\eta_p^2= .20$); however, no significant differences were found in terms of conversation quality for spatialized audio compared to mono audio, ($F_{1,23}=0.72$, $p=.40$, $\eta_p^2= .03$). Comments related to quality of the conversation included:

- *Conversations with the video turned on were a lot better.*
- *Having eye-contact and seeing other people's emotion made a huge difference and enhanced the conversation.*
- *Turning the video [on] improved the quality of the conversation.*
- *I think conversations were more effective when the monitor [video] was on because we had eye contact.*

### Engagement in the Conversation

Our participants also felt that adding video significantly increased their level of engagement in the conversation ($F_{1,23}=9.13$, $p=.006$, $\eta_p^2 = .28$); however, no significant differences were found in terms of engagement for spatialized audio compared to mono audio ($F_{1,23}=0.41$, $p=.53$, $\eta_p^2 = .02$).

When asked about the differences between the conditions, engagement was the most frequent difference raised. Comments given by the participants related to their engagement in the conversation included:

- *By adding video to the conversation I felt more engaged and willing to talk.*
- *With video [*it was*] easier to stay engaged and track the conversation. Felt accountable for joining in.*
- *I found it more engaging for me personally to actually see the people I was talking to rather than just listening to their voice on the mic.*
- *Conversations were more involved and engaging when there were videos.*
- *When I could only hear the conversation I tended to drift out and lose focus. I preferred seeing when I was having a conversation.*

A few participants however felt that they were more engaged in the conversation *without* video:

- *I prefer to be off-camera so I was more engaged in the conversations without video.*
- *I felt more engaged in the audio only conversations. But I must admit that I had no email or web distractions.*

### Knowing who is Listening in the Conversation

Our participants felt that adding video significantly increased their ability to know who was listening or paying attention during the conversation ($F_{1,23}$=18.98, p<.001, $\eta_p^2$ = .45); however, no significant differences were found in terms of listening / paying attention for spatialized audio over mono audio, ($F_{1,23}$=1.11, p=.30, $\eta_p^2$= .05).

- *Easier to pay attention (and make sure others were paying attention) with visuals.*

### Keeping track of the Conversation

A marginally significant interaction effect was found between audio type and presence of video in terms of how well participants were able to keep track of the conversation ($F_{1,23}$=3.06, p=.094, $\eta_p^2$ = .12). Further analyses revealed that when mono audio was used, the conversation was significantly easier to track when the video was present as compared to the audio only condition (Wilcoxon: z=-2.50, p=.013); however, no significant differences were found between video conditions when spatialized audio was used (Wilcoxon: z=-0.12, p=.91). Comments included:

- *When the video is off it is hard to track the conversation*
- *More difficult to have a conversation without the video to accompany it. They weren't as fluid and we would interrupt one another more often*
- *No video made it harder to show non-verbal communication. It is easier to think that pauses in the conversation mean you are not being paid attention to or that someone disagrees.*

## DISCUSSION & CONCLUSION

The results from this work strongly indicate that adding spatialized video to the audio channel significantly enhanced the distributed conversations. Participants felt that the quality of the conversation was better with the video; they felt that they were more engaged in the conversation, and they felt that it was easier to keep track of the conversation. Additional comments included:

- *Seeing other's face mattered!*
- *Nice to be able to see facial expressions and reactions.*

While the value of video supports earlier observations [4, 6], it is important to remember that the "video" used in this study is not the same as video used with most desktop video conferencing systems. Our video was spatialized, while many desktop video conferencing systems only show the speaking person. Our "spatial" video enabled the participants to see both remote colleagues and provided better gaze-awareness which may have contributed to the strong benefit we observed from the video conditions.

While most participants commented on the benefits of video, some participants did address spatialized audio:

- *Functions well with different voices coming through different speakers. No delay when people are talking simultaneously.*
- *The stereo separation really helped to identify the source.*

Interestingly, our study did not reveal any significant differences between mono audio and the spatialized audio conditions, with the exception that mono conversations without video were harder to track. This contradicts previous work which found significant benefits to spatial audio [1, 7]. It also contradicts Kilgore et al.'s result where participants preferred the spatial audio condition and perceived benefits from the spatial audio [5]. We hypothesize that the benefits from our spatialized video were stronger and therefore overshadowed any audio differences, or that our requirement that the participants in each group know each other (and their voice characteristics) diminished the benefits of spatialized audio.

Because this study did not examine non-spatial video, it is possible that regular videoconferencing with high-resolution displays could also make spatial audio redundant. Further research is needed to better understand the benefits of spatial versus non-spatial video.

The results of this work are important for the design of videoconferencing systems. Adoption remains a key barrier for teleconferencing system and users' perceptions can significantly impact their willingness to use (or not use) a system. Spatialized video clearly had a positive impact on users' experience while spatial audio benefits observed in previous work were not realized.

## REFERENCES


1. Baldis, J. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. *Proceedings of CHI 2001*, Seattle, WA, April 2001, 166-173.
2. Cutler, R., Rui, Y., Gupta, A., Cadiz, JJ, Tashev, I, He, L., Colburn, A;. Zhang, Z., Liu, Z., and Silverberg, S. Distributed Meetings: A Meeting Capture and Broadcasting System, *Proceedings of Multimedia 2002*, Juan-les-Pins, France, Dec. 2002, 503-512.
3. Ochsman, R.B. and Chapanis, A. The effects of 10 communication modes on the behavior of teams during co-operative problem solving, *International Journal of Man-Machine studies, Volume 6*, 1974, 579-619.
4. Isaacs, EA., and Tang, JC. What video can and can't do for collaboration: A case study. *Proceedings of ACM Multimedia 1993*, Anaheim, CA, Aug. 2003, 199-206.
5. Kilgore, R., Chignell, M., Smith, P. Spatialized audioconferencing: What are the benefits? IBM Conference on Centre for Advanced Studies (CASCON) 2003, Oct. 2003.
6. Sellen, A Remote Conversations: The Effects of Mediating Talk With Technology. *Human Computer Interaction, Volume 10,* 1995, 401-444.
7. Yankelovich, J., Kaplan, J., Provina, J., Wessles, M., Morris DiMicco, J. Improving audio conferencing: Are two ears better than one? *Proceedings of CSCW 2006*, Banff, Alberta, Nov. 2006, 333-342.