# AIR Conferencing: Accelerated Instant Replay for In-Meeting Multimodal Review

Kori Inkpen, Rajesh Hegde, Sasa Junuzovic, Christopher Brooks[+], John C. Tang, and Zhengyou Zhang

Microsoft Research
Redmond, WA, USA
{kori, rajeshh, johntang, zhang}@microsoft.com
sasa.junuzovic@live.com

University of Saskatchewan[+]
Saskatoon, SK, Canada
cab938@mail.usask.ca

## ABSTRACT

When people attend meetings they may miss parts of the discussion if they, for example, step out to take a phone call, go to the bathroom, or have a momentary lapse in concentration. As a result, they may need to catch up on what they missed upon returning to the meeting. Asking other attendees for a recap is often disruptive. To avoid such disruptions, we have developed an Accelerated Instant Replay (AIR) Conferencing system for videoconferencing that enables participants to privately catch up to an ongoing meeting. We explored several mechanisms where the meeting content is replayed at an accelerated rate so that the participants can catch up to the live discussion reasonably quickly.

## ACM Categories & Subject Descriptors

H5.3. Group and Organization Interfaces; H5.1. Multimedia Information Systems: Audio input/output, video; H5.2. User Interfaces.

## General Terms

Design, Experimentation, Human Factors.

## Keywords

Telepresence, videoconferencing, CSCW, collaboration, meetings, DVR, replay.

## 1. INTRODUCTION

In a perfect world, people would never arrive late to a meeting, step out of a meeting, or get distracted during a meeting. But the world is not perfect. In a recent study commissioned by Verizon Conferencing [10], 95% of participants reported missing parts of meetings, 91% reported day dreaming, 73% reported doing other work, and 39% dozed off. Additionally, many people report that they frequently multi-task during a meeting [3], especially if they are remote attendees in an audio or video conference. As a result, people often want to review material they missed.


*MM'10,* October 25–29, 2010, Firenze, Italy.


In some cases, a missed part of the meeting provides necessary context for rest of the discussion, and participants would benefit from being able to review it. The majority of the previous work in this area has focused on *post-meeting* review [4], [8], [12], [14]. In our work, we focus on enabling users to review *during* the meeting, which we refer to as *in-meeting* review. Since asking others about missed content can be disruptive to the meeting flow, we investigated mechanisms that enable participants to privately catch up. Our initial focus is on providing in-meeting review for fully-distributed, multi-way videoconferences.

Recent work by Tucker et al. [1] demonstrated the benefits of providing an audio "gist" replay of a missed part of a meeting. Their results showed that users understood a recorded meeting better and were more confident of their understanding with the replay than without. They demonstrated that compressed audio replay is beneficial when participants miss a part of the meeting.

One issue with replaying only audio is that it may not capture all of the important aspects of the missed content. For instance, facial expressions and shared workspace actions of participants will not be replayed unless video and shared workspace modification are recorded and replayed, respectively. Without facial expressions, participants reviewing the meeting may not fully understand reactions of others, and without shared workspace actions, they may not understand the verbal references to the shared workspace.

**Contributions:** To address this issue, we have built a new in-meeting replay system called Accelerated Instant Replay (AIR) Conferencing. It enables users to review meeting content in real-time during a videoconference using DVR-like features, including pause, rewind, fast-forward, and accelerated replay, as shown in Figure 1. Unlike the system presented by Tucker et al. [1], our replay mechanisms go beyond just replaying audio by supporting several replay modalities, including audio, video, shared workspace actions, and speech-to-text transcripts.

**Scope:** Our focus is on videoconferencing scenarios in which participants miss small portions of the meeting. We consider specifically the case in which participants either do not have time to review content after the meeting or understanding the information is important during the meeting.

This paper first presents relevant work for in-meeting content review. We then describe the AIR Conferencing system and discuss applicability of our work to different types of meetings and its limitations. Finally, we conclude with a discussion of the future potential of Accelerated Instant Replay (AIR) systems.

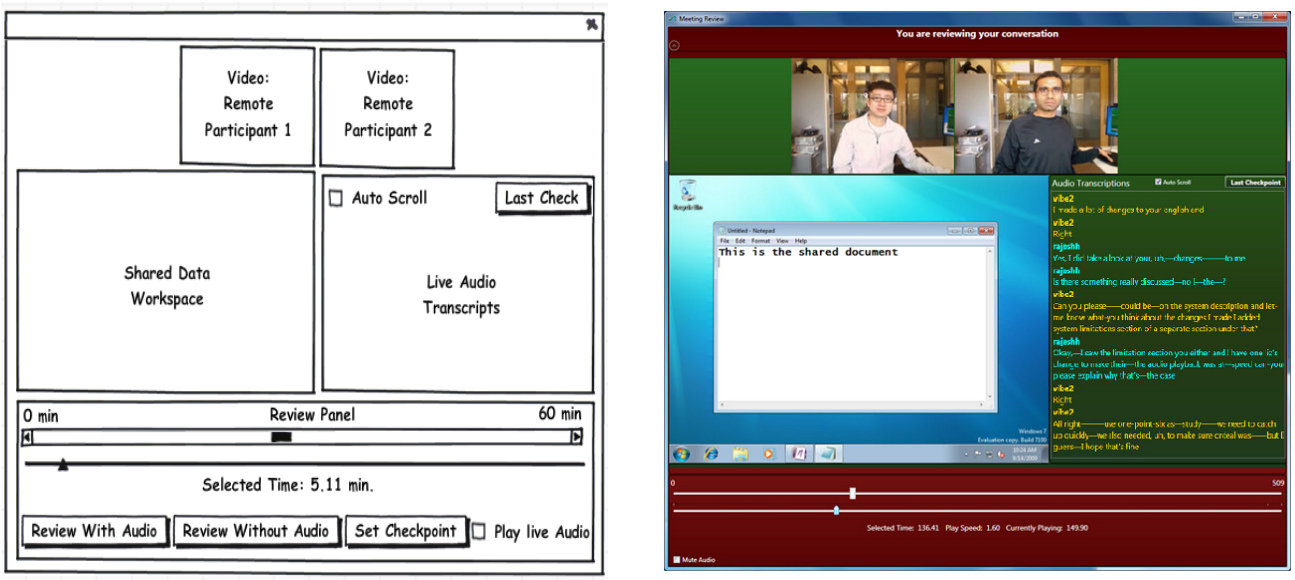


**Figure 1. AIR Conferencing Windows: (left) wireframe of the live window and (right) screenshot of the replay window.**

## 2. RELATED WORK

Post-meeting review of recorded meetings has been explored extensively in previous research. Prior work has investigated ways to facilitate automatic meeting capture [4], [8], [14], automatic content indexing [9], [16], and multimedia content replay [12] to help with such review. Previous research has also identified several key approaches for efficient multimedia playback including static summaries [6], linear compression [5], [13], and video skims [1], [2], [11], [15]. Unlike prior work on post-meeting review, we focus on in-meeting review during a live full distributed videoconference.

Little work has been done on evaluating in-meeting replay systems. Tucker et al. [1] examined an audio playback tool for meetings that uses "gisting" to compress content. This technique skims the audio providing an audio compression ratio of 2.5. While this work examined in-meeting review for pre-recorded audio, our work captures activity during a live meeting.

## 3. AIR CONFERENCING

The Accelerated Instant Replay (AIR) Conferencing system is a multi-user desktop videoconferencing system with real-time replay features. The system provides users with high quality audio and video of all participants in the videoconference, a shared workspace for data collaboration, and a real-time text transcript of the speech (see Figure 1 (left)). The main window of the system consists of four panels: 1) a video panel with videos of the remote participants; 2) a shared workspace panel; 3) an audio transcript panel; and 4) a control panel. The real-time replay component displays a new window that looks almost identical to the main window (i.e., a video panel, a shared workspace panel, and an audio transcript panel) except for different features on the control panel at the bottom (see Figure 1 (right)).

The **Video Panel** shows live videos of the remote participants.

The **Shared Workspace Panel** provides a shared desktop that all participants can interact with, but only one at a time.

The **Audio Transcript Panel** displays a real-time speech-to-text transcript of the session. The system incorporates an in-house speech recognition system to convert the participants' speech to text. The transcripts are then displayed in a text box which scrolls automatically to the most recent text. Speech-to-text conversion is done locally for each participant using the audio signal from the participant's microphone. The transcribed text is then sent to the remote participants and inserted into their transcript panels. Thus, each participant can see a flowing transcript that resembles an instant messaging conversation but preserves the interleaving turns of talk. Each participant's text is preceded with their user name and is shown in a distinct color from the other participants.

The **Control Panel** exposes several replay controls to the participants. The participants can select a start point for the replay by dragging the slider on the meeting progress timeline. They can also click on the "Set Checkpoint" button to mark the point in time at which an interruption occurred. They can then start replaying from this point when they return from their interruption.

The **Real Time Replay** component currently supports three types of replay: transcript-only replay, muted replay, and full replay.

In **transcript-only** replay mode, users can scroll through the transcript history using a scrollbar and read what was spoken in the conference. When users get interrupted during a meeting, they can click the "Set Checkpoint" button, which results in a marker appearing in the transcript history. When they return to the meeting, users click the "Last Checkpoint" button which will automatically scroll to the previous checkpoint in the transcript. When users scroll the transcript manually or click on the "Last Checkpoint" button, automatic scrolling of the transcript is disabled to afford reading the transcript history. Users can review in this mode at any time while continuing to attend to the live session in other streams (audio, video, and data).

In **muted replay** mode, the real-time replay window is shown adjacent to the main videoconferencing window and replays all media streams *except audio* (i.e., video, shared workspace, and transcript). Since the audio is muted, participants must rely on the

speech-to-text transcript to understand what was said. All of the live meeting streams, including audio, are still available, so users can continue to listen and attend to the live meeting if desired. Users can watch the replay at any speed; however, in our feedback sessions the replay speed was set to 1.6 times the normal rate. This enabled users to "*catch up*" to the live meeting in reasonable time.

In **full replay** mode, the real-time replay window is also shown adjacent to the main videoconferencing window, but it replays *all the streams* (audio, video, shared workspace, and transcript). As with the other two replay modes, users can begin replay from an arbitrary point in the past of from the previous checkpoint. All of the live meeting streams (except audio) are available during the replay, so users can attend to the live meeting while replaying the past; however, they must rely on the live transcript to understand what is being said. As in the muted replay mode, the replay speed was set to 1.6 times the normal rate.

To solve audio pitch issues that arise when audio is played back at an accelerated rate, we used an in-house audio speed-up technology [7] that employs pitch correction and silence adjustment techniques. As mentioned above, we replay content at the rate of 1.6 times the normal speed. We chose this rate over 1.4, which was used previously [1], and 2.0, which has been shown to be on the upper end of what users can understand [1], [18]. Through our own pilot testing we felt that 1.6 is a good compromise between speed and understandability.

The AIR system supports up to seven simultaneous users in a single conference. This limitation arises from computational requirements of video encoding and bandwidth requirements for sending media streams over the network (Tested on 3GHz dual core machines over 100Mbps LAN network).

# 4. PRELIMINARY USER FEEDBACK

Eighteen participants, seventeen male and one female, between the ages of 24 and 45 were recruited to try out the system. The participants all had a technical background and were comfortable with technology. Participants were recruited in groups of three and all members of a group knew each other well.

The participants were asked to fill out a short background questionnaire before being introduced to the system. Participants were then given a ten-minute training session on the AIR Conferencing system and then asked to participate in a mock "status update" type of meeting, where each person in a meeting was asked to give a short presentation.

During each presentation, the participants were interrupted twice and asked to step outside the office for 90 seconds. This ensured that some content from the presentations was missed. After the interruption the participants were instructed to catch up on what they missed using one of the three replay modes.

# 5. RESULTS

When asked whether they felt it would be useful to replay what they missed during a videoconference ten participants indicated yes, seven indicated maybe, and one indicated no. As one participant explained, "*Often you miss critical conversations when you step out or are interrupted during a meeting and then you try to play catch up during the rest of the meeting. Getting to know what was covered and who said it and the body language would put me back into the meeting very quickly*."

While many of the participants felt that it would be beneficial to "*get context and avoid interrupting the meeting with questions that had already been covered,*" for a number of participants, the answer ultimately depended on the context of the meeting, how much they missed, and the type of replay mechanism. As one participant explained, "*It would depend mostly upon the importance of the meeting, followed by the duration of how much I missed, and finally, on how discreetly I could review the video*".

Some participants expressed concern about whether the replay would be disruptive to the rest of the meeting – "*if it can be done discreetly, then all the better.*" One participant voiced a concern that replay may cause them to miss even more of the meeting, "*I wouldn't want to miss more information (while) reviewing missed parts.*" Another participant mentioned that "*reviewing the video may force me to stay behind, but "maybe" if it looks like an unimportant or uninteresting (part of the) conversation is taking place and that gives me a window to catch up*".

After using the AIR Conferencing system, eleven participants strongly agreed that it would be useful to use a replay system like AIR, six somewhat agreed, and one somewhat disagreed.

While most participants (16/18) preferred the full-replay condition, two favored the transcript-only condition. Participants that chose the full replay condition commented that "*it was fast, easy to concentrate and auto catch up*" and "*I can listen to audio while watching the live slide show and transcript.*" The participants who preferred the transcript-only replay condition commented that "*you can listen to current conversations and read transcripts*" and "*it is easier to multi-task.*"

One explanation for why only a few users favored the transcript-only condition is that, despite the fact that we trained the speech-to-text system for each user, the quality of the transcript was problematic for many users. Four participants explicitly commented that they preferred the full replay condition because the "*transcript quality was low*" in the other conditions.

# 6. IMPLEMENTATION CHALLENGES

We encountered several implementation challenges for in-meeting replay. One issue related to accelerated audio playback was handling the case when replay audio catches up to the live audio. Traditional streaming audio technologies do not allow management of a live, real-time audio channel and an accelerated catch up audio channel. In our system, we stored up to 60 minutes of audio in memory using a circular buffer, which enabled us to continuously write live audio to the end of the buffer while also allowing the user to review playback from any point in the buffer.

Another challenge was to synchronize other multimedia streams, such as video, data, and transcription, with the accelerated audio replay stream. This was particularly important for the transcription stream since they had to be interleaved in the same order as the participants' speech. Furthermore, each participant's computer is independently generating audio and transcript data for that participant. For the sake of simplicity, we used local participants' audio timings as the reference which in turn used the local system clock. Since all the computers we used synchronized their times to a network time server, this approach worked well. In a few cases the audio transcription engine took longer to transcribe some participants' speech which caused some of the transcriptions to be out of order with respect to the other streams.

Although we used an in-house state of the art speech-to-text algorithm in our system (with 30 minutes of training per participant), the speech-to-text conversion was not very accurate. Several issues make this matter potentially worse in terms of future use of such systems for in-meeting review. First, for many participants, there may be no training data in which case the accuracy of the speech-to-text system would be even lower than what we observed. Second, the frequent use of jargon and abbreviations are difficult for speech-to-text systems to understand. Finally, speech-to-text engines have difficulties with accents. These issues must be resolved to improve the usefulness of speech-to-text systems for in-meeting review.

# 7. CONCLUDING REMARKS AND FUTURE WORK

This work makes several significant contributions. First, we designed and built a new in-meeting replay system that goes beyond just replaying audio by incorporating audio, video, shared workspace actions, and a speech-to-text transcript into an accelerated playback review. Preliminary user feedback sessions were run using our prototype during a series of live videoconference meetings. The results show that users saw value in an accelerated instant replay system for in-meeting review.

Most of the users in our study had previous experience with both videoconferencing systems and DVR systems, and see benefits to both. Given that people take advantage of DVRs when they miss or do not understand something in a television program, it is plausible that this behavior could translate to videoconferencing. When asked about this possibility, most of the users in our study strongly felt that these features would be useful.

When developing an accelerated instant replay system, several design considerations must be taken into account. Feedback from our research indicates that there is a cost-benefit tradeoff. The missed information needs to be important enough to warrant review and the system needs to be easy to use; otherwise, people may not use it. Users are also concerned about disrupting or missing more of the meeting while catching up. Thus, the system should be designed so that the replay process is seamless and does not detract from the meeting or disrupt the flow of the meeting (both for the person catching up and other meeting participants).

We have explored three replay mechanisms in our AIR system; however, we feel that we have barely scratched the surface of in-meeting replay support. We plan to explore the potential of each of the different components (video, audio, transcript, and shared documents) to better understand how they can be combined to provide maximum benefit. We also plan to study how well users can attend to both the past and the present at the same time.

# 8. ACKNOWLEDGMENTS

We would like to thank Kit Thambiratnam and Frank Seide from the MSR Asia Speech group for their help with the speech-to-text component. We would also like to thank Christian Huitema and our MSR colleagues from CCS, VIBE, and Connect. Finally, we would like thank the people who took part in the feedback session.

# 9. REFERENCES

[1] Christel, M., Winkler, D., Taylor, R., and Smith, M. 1998. Evolving video skims into useful multimedia abstractions. *ACM CHI*. 171-178.

[2] Christel, M. 2006. Evaluation and user studies with respect to video summarization and browsing. *Symposium on Eletronic Imaging 6073*.

[3] Chudoba, K.M., Watson-Manheim, M.B., Lee, C.S., and Crowston, K. 2005. Meet me in cyberspace: meetings in the distributed work environment. *Academy of Management Conference*.

[4] Cutler, R., Rui, Y., Gupta, A., Cadiz, JJ, Tashev, I., He, L. Colburn, A. Zhang, Z., Liu, Z., and Silverberg, S. 2002. Distributed meetings: a meeting capture and broadcasting system. *ACM Multimedia*. 503-512.

[5] Dietz., P. H. and Yerazunis, W.S. 2001. Real-time audio buffering for telephone applications. *ACM UIST*. 93-94.

[6] He, L., Sanocki, E., Gupta, A., and Grudin, J. 1999. Auto-summarization of audio-video presentations. *ACM Multimedia*. 489- 498.

[7] He, L. and Gupta, A. 2001. Exploring benefits of non-linear time compression. ACM MULTIMEDIA '01, vol. 9. 382-391.

[8] Jain, R., Kim, P., and Li, Z. 2003. Experiential Meeting System. *ACM SIGMM workshop on Experiential Telepresence. 1-12.*

[9] Kazman, R., Al-Halimi, R., Hunt, W., and Mantei, M. 1996. Four paradigms for indexing video conferences. *IEEE Multimedia.* 3, 1. 63-73.

[10] Meetings in America V: meeting of the minds. 2003. https://e-meetings.verizonbusiness.com/global/en/meetingsinamerica/uswhitepaper.php.

[11] Money, A. G. and Agius, H. 2008. Video summarization: a conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Recognition*. 19. 121-143.

[12] Moran, T., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., and Zellweger, P. 1997. I'll get that off the audio: a case study of salvaging multimedia meeting records. *ACM CHI*. 202-209.

[13] Omoigui, N., He, L., Gupta, A., Grudin, J., and Sanocki, E. 1999. Time-compression: systems concerns, usage, and benefits. ACM CHI. 136-143.

[14] Ranjan, A., Birnholtz, J. and Balakrishnan, R. 2008. Improving meeting capture by applying television production principles with audio and motion detection. *ACM CHI*. 227-236.

[15] Smith, M. and Kanade, T. 1995. Video skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-95-186.

[16] Tucker, S. and Whittaker, S. 2004. Accessing Multimodal Meeting Data: Systems, Problems, and Possibilities. *Workshop on Multimodal Interaction and Related Machine Learning Algorithms.*

[17] Tucker, S., Bergman, O., Ramoorthy, A., and Whittaker, S. 2010. Catchup: a useful application of time-travel in meetings. *ACM CSCW*. 99-102.

[18] Wildemuth, B., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., and Gruss, R. 2003. How fast is too fast? Evaluating fast forward surrogates for digital video. ACM/IEEE-CS Conference on Digital Libraries. 221-230.