What did I miss? In-Meeting Review using Multimodal Accelerated Instant Replay (AIR) Conferencing

Sasa Junuzovic[1], Kori Inkpen[1], Rajesh Hegde[2], Zhengyou Zhang[2], John Tang[1], and Christopher Brooks[3]

Connect[1] and CCS[2] Groups
Microsoft Research, Redmond, WA
{sasajun, kori, rajeshh, zhang, johntang}@microsoft.com

Department of Computer Science[3]
University of Saskatchewan, Saskatoon, SK
cab938@mail.usask.ca

ABSTRACT

People sometimes miss small parts of meetings and need to quickly catch up without disrupting the rest of the meeting. We developed an Accelerated Instant Replay (AIR) Conferencing system for videoconferencing that enables users to catch up on missed content while the meeting is ongoing. AIR can replay parts of the conference using four different modalities: audio, video, conversation transcript, and shared workspace. We performed two studies to evaluate the system. The first study explored the benefit of AIR catch-up during a live meeting. The results showed that when the full videoconference was reviewed (i.e., all four modalities) at an accelerated rate, users were able to correctly recall a similar amount of information as when listening live. To better understand the benefit of full review, a follow-up study more closely examined the benefits of each of the individual modalities. The results show that users (a) preferred using audio along with any other modality to using audio alone, (b) were most confident and performed best when audio was reviewed with all other modalities, (c) compared to audio-only, had better recall of facts and explanations when reviewing audio together with the shared workspace and transcript modalities, respectively, and (d) performed similarly with audio-only and audio with video review.

Author Keywords

Telepresence, videoconferencing, CSCW, meetings, DVR, review, audio, video, shared workspace, transcript.

ACM Classification Keywords

H.5.1. Multimedia Information Systems; H.5.2. User Interfaces; H.5.3. Group and Organization Interfaces.

General Terms

Experimentation, human factors.

INTRODUCTION

Many people participate in meetings as part of their daily work. They often miss parts of meetings because of interruptions, distractions, or multitasking [7]. As a result, they may need to review the material they missed. The majority of the previous work in the area of meeting review has focused on *post-meeting review* [2,9,12]. Post-meeting review is beneficial when the missed segments are not critical for the rest of the discussion and can be reviewed after the meeting has finished. However, in other cases, the missed content provides necessary context for rest of the discussion, and participants would benefit from being able to review it during the meeting. 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, we investigated mechanisms that enable participants to privately catch up. Our initial focus is on providing in-meeting review for fully-distributed videoconference meetings. Additionally, our focus is on scenarios in which participants miss small portions (e.g., a minute or two) of the meeting.

A key requirement for in-meeting review is that it should enable users to review past content without missing new content being generated in the live discussion. One approach is to present only past information during review periods but present it in such a way that users can catch up to the live discussion. Recent work by Tucker et al. [14] demonstrated the benefits an audio-gisting technique which provides a compressed version of the missed audio. Their results showed that users understood the meeting better and were more confident in their understanding when they were able to review the missed content using audio-gisting.

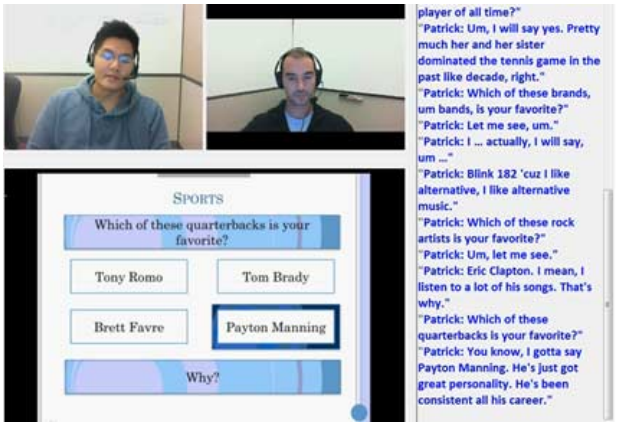

Figure 1. The AIR Conferencing window

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Our work investigates the potential of accelerated *multimodal* review for in-meeting scenarios. We built a new system that enables users to review content in real-time during an ongoing meeting called Accelerated Instant Replay (AIR) Conferencing [5], shown in Figure 1. AIR incorporates DVR-like features, including pause, rewind, and accelerated review, to support several catch-up modalities including audio, video, shared workspace actions, and conversation transcript. We evaluated the benefit of AIR through two user studies. The first study explored the benefits of AIR catch-up during a live meeting. The results showed that when reviewing the full videoconference (i.e., all four modalities) at an accelerated rate, users were able to correctly recall a similar amount of information as when listening live. To better understand the benefit of full review, a follow up study more closely examined the benefits of each of the individual modalities. The results show that users (a) preferred using audio along with any other modality to using audio alone (b) were most confident and performed best when audio was reviewed with all other modalities, (c) compared to audio-only, had better recall of facts and explanations when reviewing audio together with the shared workspace and transcript modalities, respectively, and (d) performed similarly with audio-only and audio with video review.

Next, we present relevant work for in-meeting review. We then describe the AIR Conferencing system. Then, we describe the results of our initial study that evaluated the system in a three-way distributed meeting. Following this, we describe the results of our second, more in-depth study that more closely examined the benefits of each of the individual modalities. Finally, we conclude with discussions, the future potential of in-meeting catch-up systems, and the insights we gained from this work.

RELATED WORK

Providing users with the ability to review recorded meetings has been explored extensively in previous research. Prior work has investigated ways to facilitate automatic meeting capture [2,12], ways to automatically index meetings [6,13], and methods to replay multimedia content [9]. Compared to these approaches, our work focuses on fully distributed videoconferences and addresses techniques to review multimedia content.

Previous research has identified several key approaches for efficient multimedia playback including static summaries [4], linear compression [3], and video skims [1,8,10]. Static summary systems convert video segments into less rich modalities, such as a textual summary or a single image [4]. While they can be helpful for reviewing large amounts of content quickly, they reduce the fidelity of information by removing temporal aspects and serve as a lossy conclusion of what was presented. These issues can be solved by compressing content instead of changing its modality, which is the approach taken by linear compression systems. Linear compression systems compress multimedia content by dropping audio and video frames in a systematic fashion [3]. However, dropping frames may degrade the replay experience. While the resulting quality problems with video are difficult to solve, audio quality issues, such as pitch shift, can be corrected [10]. Experimental results suggest that a compression ratio of 2 is reasonable [11,15] for audio. The low computational overhead and the lossless features of this approach make it tractable as a method of catching up in a live videoconference. Linear compression, however, is not content driven. In many cases, certain sequences of audio and video are more important than others, and it would be useful to remove segments (e.g., dead air between speakers in an interview) or overlap segments (e.g. audio over a collage of video) to allocate more time to the important parts. Video skimming [1,8] is a technique that takes into account the context of content to create abbreviated content-driven summaries.

More recently Tucker et al. [14] introduced an in-meeting audio-catch-up mechanism that uses a skimming technique called gisting. Their tool summarizes an audio recording of a meeting using a three-stage post-production process: 1) an automated speech-to-text system generates a transcript; 2) a summarization scheme extracts the important information, the gist; and 3) an audio file containing only the gist is generated. Their system uses a compression ratio of 2.5.

Contrasting our catch-up approach to Tucker et al.'s gisting technique, we find that one of the key benefits of gisting is that it can often compress audio at higher rates than those provided by typical acceleration algorithms. However, the quality of the resulting gist depends on the accuracy of the mechanisms used to create the gist. Current state of the art speech-to-text systems typically require several hours of training for each user to gain reasonable levels of accuracy. Also, gisting can result in the loss of information because the gist summarizes most but not all content.

AIR CONFERENCING

The AIR (Accelerated Instant Replay) Conferencing system [5] is a multi-user desktop videoconferencing system with in-meeting review features. The videoconferencing portion of the system provides users with high quality audio and video of all participants, a shared workspace for data collaboration, and a real-time text transcript of the audio.

When observing the live content (i.e., when not using the in-meeting review feature), users interact with AIR through a window containing four panels shown in Figure 2 (left). The Video panel shows live video feeds for all remote users. The Shared Workspace panel provides a shared workspace that any user in the conference can interact with. The Transcript panel displays the conversation transcript generated by an automatic speech recognizer. Each user's text is preceded with the user's name and is shown in a different color than the text of other users. The Review panel contains review controls to mark catch-up sections and start in-meeting review.

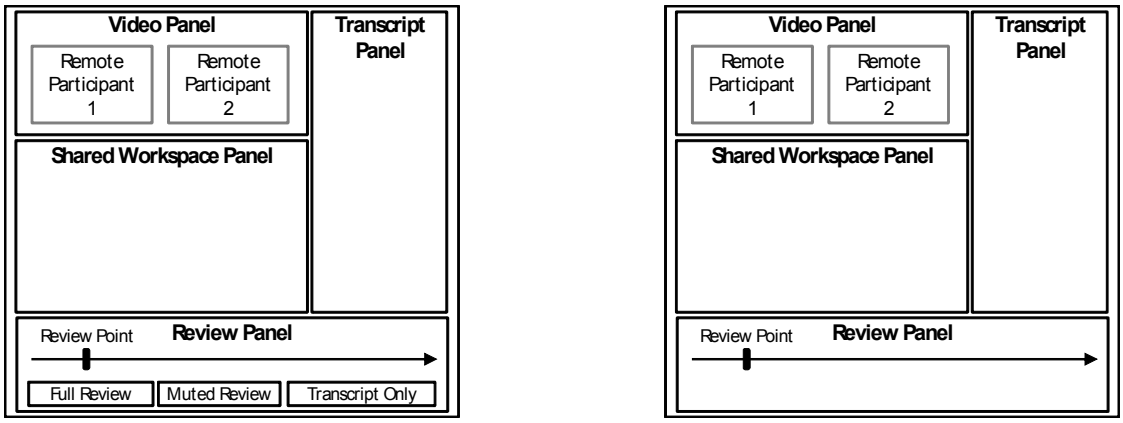

Figure 2. AIR Conferencing windows: (left) the live window and (right) review window

When using the in-meeting review feature, users interact with AIR using a separate window, shown in Figure 2 (right), that contains the same four panels as those in the live window with minimal differences. In particular, the Control panel displays a timeline allowing users to select a section of the conference to be reviewed. The other three panels display the reviewed content at an accelerated rate instead of the current meeting discussion.

Both the live and review windows are visible during catch-up sessions, enabling users to monitor the live meeting while simultaneously reviewing content from an earlier point in time. Users can choose whether to listen to the live meeting or past audio, but not both. By default, the playback is accelerated to a rate of 1.6 times normal to enable catching up to the live meeting. We solved the audio pitch and intelligibility issues that arise when audio is played back at an accelerated rate by using an audio speed-up technology that employs pitch correction and silence adjustment techniques to make sure the audio sounds meaningful even when played back at a faster rate [10].

USER STUDY #1

We performed a user study to get a preliminary evaluation of the AIR system. Our goals were two-fold. First, we wanted to measure general interest in a DVR-like approach for in-meeting review. Second, we wanted to evaluate the benefit of review mechanisms that go beyond just audio.

Participants and Procedure

We recruited eighteen participants (one female) between the ages of 24 and 45 (median 30), from a large software company. The participants all had a technical background, were comfortable with technology, and most (twelve) considered themselves to be early adopters of technology. Participants were recruited in groups of three. All members of a group knew each other well (for at least six months).

We first conducted 30 minute interviews with each participant to gather background data, measure interest in a system like AIR, and train the speech-to-text system used by our prototype. Next, the groups were brought into our research facility to take part in a distributed group meeting. The meeting took place in three adjacent usability lab rooms. Upon arrival the participants were given a brief introduction to the study and then asked to fill out a short background questionnaire. Participants were then given a ten-minute training session on the AIR Conferencing system, after which they began the main study task.

The main study task was designed to mimic a "status update" meeting, where each person is asked to give a short presentation on their project. These types of meetings are common in business environments. Instead of requiring our participants to prepare a presentation, we provided them with a presentation that prompted them for their preferences in a number of areas (e.g., technology, media, social) through multiple choice questions. Each participant was given twelve minutes to answer as many questions as possible (maximum seventy-six). For example, one question was "Which of these web browsers is your favorite?" The participant (a) read each question out loud, (b) picked one of the choices, and (c) gave a short explanation. The other group members were asked to remember answers and explanations. Each group member answered the same set of questions but in a different order than the other group members.

We wanted to simulate interruptions during real meetings. Thus, during each presentation, the participants experienced two interruptions where they left the experiment room and stayed out for 90 seconds before rejoining the meeting. This ensured that content from the live presentation was missed. After an interruption, the participants were instructed to review the portion of the meeting that they missed using one of three catch-up techniques.

At the end of the meeting, each participant completed two quizzes. Each quiz repeated the multiple choice questions answered during one of the presentations the participant observed. Each question also asked for the explanation given by the presenter. The question order in the quiz corresponded to that of the presentation.

Experimental Design

We explored three different review conditions in this study: *transcript-only*, *muted-review*, and *full-review*. In the transcript-only condition, users manually scrolled the speech-to-text window so that they could see a transcript of the conversation that they missed. No separate review window was launched and audio from the live meeting was still played. In the muted-review condition, a muted version of the video conference was reviewed in a separate window, including video, shared workspace (i.e., the presentation), and transcript. The live meeting was still visible in the live window, and its audio was still played. In the full-review condition, a full version of the video conference was reviewed in a separate window, including audio, video, shared workspace, and transcript. The live meeting was still visible in the live window, but its audio was muted.

In the muted and full-review conditions, content was reviewed at a rate of 1.6 times the normal speed. We chose this rate over 1.4 times, which was used previously [14], and 2.0 times the normal speed, which has been shown to be on the upper end of what users can understand [14,15]. Through our own pilot testing, we felt that this rate was a good compromise between speed and understandability.

As mentioned above, each participant gave one presentation for a total of three presentations per group. Each non-presenting participant was interrupted twice during a presentation. After an interruption, the participant used one of the catch-up conditions to catch up. All conditions and orders were counterbalanced to create eight orderings.

RESULTS OF USER STUDY #1

Background Interviews[1]

The background interview gathered information on users' previous experiences with missing parts of meetings, videoconferencing systems, DVR systems, and their interest in having a real-time review option during meetings. All of the participants reported sometimes missing parts of meetings, typically for less than five minutes. While they employ various strategies to review missed content, none of the strategies included viewing recordings of other participants or whiteboards. Thus, some content, such as facial expressions, gestures, and whiteboard data, are apparently never reviewed during meetings. Moreover, no participant brought up the idea of recording a meeting for in-meeting review, even though many of them use DVRs to pause, replay, and fast-forward TV content to handle scheduling conflicts, breaks, and interruptions.

When asked whether they felt it would be useful to review what they missed during a videoconference, ten participants indicated yes, seven indicated maybe, and only one indicated no. 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 catch-up mechanism. 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 were concerned that the review might be disruptive to the rest of the meeting, "*If it can be done discreetly, then all the better*." Also, there was concern that if they spent time catching up on missed information, they would end up missing 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*." One participant expressed desire for an accelerated review feature, "*If it was short and especially if I had a way of speeding up the content or seeing a transcript to help me multi-task and also not miss any (new) content.*"

These interviews suggest that users could benefit from a DVR-like in-meeting review system that allowed them to review missed content and catch up to the live discussion quickly.

Playback Options

We analyzed the participants' scores on the quizzes from the group session to assess the effectiveness of the catch-up techniques on participants' recall. The system crashed for one participant so we report on data from seventeen participants in this section. We calculated a baseline recall score for each participant based on the percentage of correct answers they had for the parts of the presentations they viewed live. We compared the participants' scores for the baseline, as well as the transcript-only, muted-review, and full-review techniques. We also examined these results for

Recall	Facts	Explanations
Baseline (Live)	78%[1]	43%[1]
Transcript Only	45%	16%
Muted Review	40%	13%
Full Review	80%[1]	49%[1]

Table 1. Percentage of correct quiz answers. [1]Recall was significantly higher in the baseline and full-review conditions compared to the transcript-only and muted-review conditions ($p<.05$). There were no significant differences between the full-review and baseline conditions, $p>.05$).

[1] Some of the results from the background interviews were presented in a poster at ACM Multimedia 2010 [5].

both the multiple choice answers (facts) as well as the short answers (explanations). The percentage correct for each category is shown in Table 1.

A 2-answer type {fact, explanation} x 4-experimental condition {baseline, transcript-only, muted-review, full-review} repeated measures ANOVA revealed significant main effects for answer type ($F_{1,16}$=74.8, $p<0.001$) and experimental condition ($F_{3,48}$=22.8, $p<0.001$), but no significant interaction effect ($F_{3,48}$=0.6, $p=0.63$).

Pair-wise post-hoc analyses of the experimental condition results (with Bonferroni corrections) revealed that participants' recall using the full-review condition was not significantly different than their baseline recall ($p=1.0$). However, recall in the transcript-only and muted-review conditions were significantly worse than both the full-review condition and the baseline recall ($p<0.01$).

Examining the results by answer type revealed that the participants performed significantly worse on the explanation answers than the fact answers, and this result was consistent across the experimental conditions ($p<.05$).

Final Questionnaire Data

After completing the main task and the quizzes, the participants were asked whether it would be useful to use a system like AIR to catch up on what they missing during a meeting. A 5-point scale (1 = strongly disagree and 5 = strongly agree) was used. Eleven participants strongly agreed that it would be useful to use a catch-up system like AIR, six somewhat agreed, and one somewhat disagreed.

Participants were also asked to report which of the three catch-up mechanisms they preferred. Significantly more participants choose the full-review condition (16) than the other two conditions ($\chi 2$=25.33, $p<0.001$). These participants commented that:

- *"It was fast, easy to concentrate and auto catch-up"*
- *"Easier to playback and comprehend"*
- *"Ignore what was live as I had full fidelity"*
- *"Leads most smoothly into rejoining live"*
- *"The only one that let me digest information"*
- *"Seems like the only way to catch up in a focused way"*
- *"I feel like I can cheat in time with fast forward"*
- *"I can listen to audio while watching the live slide show and transcript"*

The other two participants chose the transcript-only condition because "*You can listen to current conversations and read transcripts*," and "*It is easier to multi-task*."

New Questions

Our initial study raised three new questions. First, the results indicated that when all four modalities are used for in-meeting review, users were able to correctly recall the same amount of information as when listening live. As described earlier, Tucker et al. [14] found that audio-only review is also a helpful catch-up mechanism. An important question is whether the additional modalities (video, shared workspace, and transcript) have additional benefits.

Second, despite the fact that we trained the speech-to-text system for each user, the quality of the conversation transcript was poor for many users. This effect likely had a significant impact on users' performance and preference. Four participants explicitly commented that they preferred the full-review condition because the "*transcript quality was low*" in the other conditions. We asked participants to re-rank the conditions assuming perfect transcription (Group 1 was not asked to answer this question, so n=15 for this result). Assuming perfect transcripts, there were no significant differences between the conditions ($\chi 2$=2.80, $p=0.247$). The idea of a perfect transcript swayed some users away from the full-review to transcript-only and muted review: seven preferred transcript-only; six wanted full-review; and two desired muted-review. Therefore, it is important to reevaluate the benefit of transcript-only review when the speech-to-text system is perfect.

Third, the fact that users swayed away from full-review when the transcript is perfect indicated that they believe that they can use the transcript to catch up on past information while simultaneously listening to the live audio. To investigate further, we asked the participants to indicate whether it was difficult to attend to both the past and the present at the same time on a five point scale (strongly disagree to strongly agree). Fourteen people strongly agreed (six) or agreed (eight) that attending both the past and the present at the same time was difficult. Two participants found it neither easy nor difficult, and two found it somewhat easy. The debrief interviews revealed that while the participants had different preferences for whether or not they would like to hear live audio and read the past transcript or vice versa, one constant across all participants was that if the speech-to-text engine generated more accurate transcripts, it would have been easier to pay attention to both past and present at the same time. As it were, reading and understanding the transcript made it difficult to listen to audio concurrently.

Next, we describe a second user study of in-meeting review mechanisms that addressed the role of audio relative to other catch-up modalities and accuracy of the transcript. We leave the issue of divided attention for future work.

USER STUDY #2

Given the success of the AIR full-review configuration from the first study, we wanted to more closely examine the benefits of each of the individual modalities: audio, video, shared workspace, and conversation transcript.

Participants and Procedure

58 participants (25 female) between the ages of 18 and 60 (median of 39.5) were recruited for this study. Upon arrival, participants were introduced to the concept of accelerated

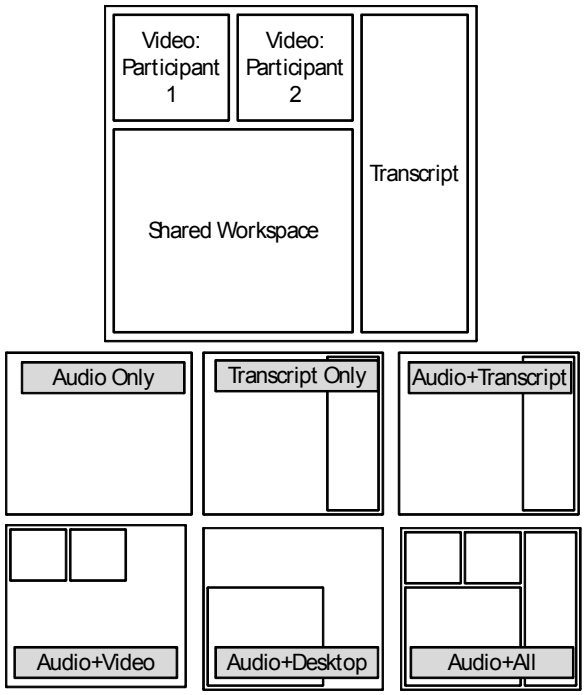

Figure 3. The mockup AIR system: (top) live window and (bottom) audio only, transcript only, audio+transcript, audio+video, audio+desktop, and audio+all review.

instant review and were told that they would be attending two, three-way distributed meetings using an in-house videoconferencing tool. They were also told that they would be interrupted and asked to catch-up on what they missed using two different catch-up mechanisms.

The simulated status meeting task from the first study was re-used in this study, modified slightly to have two presenters per meeting instead of one. To maintain consistency, two meetings were pre-recorded using two actors who answered questions from the shared presentation. The participants took on the role of the third person in the meeting (a passive attendee). For both meetings, the presentation, complete with audio, videos, and shared workspace, were recorded. The audio was also manually transcribed and synchronized with the audio and video to ensure a perfect transcript. The meetings were then played back using a system that matched the look and feel of AIR (see top image in Figure 3). Because of the divided attention challenges reported in the first study, this version of AIR showed either the review or the live window, but not both, and automatically switched between them.

Each meeting consisted of 40 questions, where the presenters alternated every five questions. Additionally, the presenters clicked on their answers in the shared workspace to provide visual confirmation of their choice. The first meeting was 7.5 minutes long and the second meeting was 8.25 minutes long. Each meeting had a 1 minute initial live portion, a 1.5 minute interruption, a 2.5 minute AIR catch-up portion, and a second live session 2.5 or 3.25 minutes long. The lengths were different because the actors took longer to answer questions in the second meeting.

After each meeting, the participants were asked to complete a quiz which contained all questions from the meeting but in a random order. The participants had to recall (a) the answer to the question, (b) the identity of the presenter who answered it, and (c) the explanation given for the answer. Following the quiz, the participants completed a questionnaire regarding their confidence in their answers and their experience during the live and review parts of the meeting. At the end of the session participants also completed a free-form questionnaire regarding their preferences for the catch-up mechanisms they used.

Experimental Design

Five different configurations of review modalities were examined in this study: 1) *audio only*; 2) *transcript only*; 3) *audio+transcript*; 4) *audio+video*; 5) *audio+workspace*; 6) *audio+all* (video, transcripts and shared workspace). These six conditions are shown in the six smaller images in Figure 3. We refer to conditions three through six as the *enhanced-audio review conditions*. The audio-only condition was used as a baseline for all 58 participants. The remaining five conditions were evaluated as a between subjects factor with each participant completing the task using one of these modalities (in addition to the audio-only condition). Each of the enhanced-audio conditions had twelve participants, while the transcript-only condition had ten. Condition and meeting order were counterbalanced across participants.

RESULTS OF USER STUDY #2

The main goal of this study was to explore whether video, shared workspace, and transcript added value over an audio-only catch-up. We first analyzed results for the four enhanced-audio conditions (audio+all, audio+workspace, audio+transcript, and audio+video) compared to audio-only and live (see Tables 2, 3, and 4), followed by a separate analysis of the transcript-only condition (see Table 5).

For the recall results, we analyzed participants' answers for the multiple choice questions (*facts*), the short answers (*explanations*), and identification of the person answering the question (*identifications*). We calculated recall scores for each participant based on the percentage of correct answers they had for the questions they viewed live, the catch-up questions in the audio-only condition, and the catch-up questions in the enhanced-audio or transcript-only conditions. For the preference and confidence ratings, we analyzed ratings from the post-meeting questionnaires which were based on a 7-point scale where 7 was extremely confident (see Table 4).

Enhanced Audio

Of the 48 participants who participated in the enhanced-audio review conditions, significantly more (40) preferred the enhanced-audio over the audio-only condition (χ^2=33.3, $p<.001$). This result was also supported by participants'

Recall	Fact	Explanation	Identification
Baseline (Live)	93%	77%	84%
Audio Only	74%	50%	61%
Enhanced Audio	83%[1]	62%[1]	72%[1]

Table 4. Percentage of correct answers for facts, explanations, and identifications. [1]Enhanced audio was sig. higher than audio only, but sig. lower than live, p<.05.

Recall	Fact	Explanation	Identification
Baseline (Live)	93%[1]	77%[1]	84%[1]
Audio Only	74%[2]	50%[2]	61%[2]
Audio+All	89%[1]	70%[1]	81%[1]
Audio+Workspace	90%[1]	65%[2]	72%[2]
Audio+Transcript	83%	67%[3]	66%
Audio+Video	70%[2]	48%[2]	68%[2]

Table 4. Percentage of correct answers for the enhanced-audio conditions. [1]Not statistically different from live but significantly better than audio only. [2]Not significantly different than audio only. [3]Recall of explanations significantly better than audio only, but significantly lower than live. Significance was measured as *p*<.05.

Confidence	Overall	Fact	Explanation	Ident.
Audio Only	2.7	3.8	3.4	3.7
Audio+All	[1]5.4	[1]5.1	[1]5.4	[1]5.9
Audio+Workspace	3.8	3.3	3.5	4.8
Audio+Transcript	4.9	3.6	4.5	4.6
Audio+Video	3.3	4.9	3.6	4.4

Table 4. Average responses to confidence level questions for being caught-up on all information during the catch-up phase (overall), and correctly answering multiple choice questions, explanations and identifications.

ratings of how much they liked each catch-up mechanism. They liked the enhanced-audio conditions significantly more (mean 4.9) than the audio-only condition (mean 3.1), ($F_{1,28}$=26.49, *p<.001*). Examining each of the enhanced-audio conditions separately, we found that all conditions were significantly preferred over the audio-only condition (*p<.05*) except for the audio+transcript condition (*p=.06*).

Recall results were analyzed using a mixed repeated measures ANOVA, with two within subject variables: *session* {live, audio, enhanced audio} and *question type* {fact, explanation, identification}; and two between subjects variables: *condition* {audio+all, audio+workspace, audio+transcript, audio+video} and *gender*. Bonferroni corrections were used for all post-hoc analyses. The results are shown in Table 2 and 3. A significant interaction effect was found for session and question type ($F_{4,152}$=3.53, *p*=.009), so we examined each question type separately. No significant effects were found for gender ($F_{1,38}$=2.38, *p*=.13).

Enhanced Audio: Significant main effects of session were found for all question types (facts: $F_{2,88}$=48.42, *p*<.001; explanations: $F_{2,84}$=61.08, *p*<.001; and identification: $F_{2,88}$=48.52, *p*<.001). Participants had significantly higher recall in the enhanced-audio session than the audio-only session but significantly lower recall when live (*p*<.05) as shown in Table 2.

Audio+All: For all question types, recall was significantly higher in the audio+all condition than the audio-only condition (*p*<.05), but not significantly different than live (*p*>.05) as shown in Table 3.

Audio+Workspace: For facts, recall was significantly higher in the audio+workspace condition than audio-only (*p*<.05) but not significantly different than live (*p*>.05). For explanations and identification, recall in the audio+workspace condition was significantly lower than live (*p*<.05) and not significantly different than the audio-only condition (*p*>.05).

Audio+Transcript: The results were mixed for the audio+transcript condition. For facts and identifications, the audio+transcript condition was not significantly different than either the live or the audio-only conditions (*p*>.05). For explanations however, the audio+transcript condition had significantly higher recall than the audio-only (*p*<.05), but significantly lower than the live condition (*p*>.05).

Audio+Video: For all question types, recall was significantly lower in the audio+video condition than the live condition (*p*<.05) but not significantly different than the audio-only condition (*p*>.05).

During the post-meeting questionnaire, we asked participants to rate their confidence in the accuracy of their answers overall and for facts, explanations, and identifications separately (see Table 4). Wilcoxon signed ranks tests revealed that for all confidence ratings, participants were significantly more confident in their answers in the audio+all condition than the audio-only condition (*p*<.0125). None of the other conditions had significantly different confidence ratings compared to audio-only.

Feedback from participants on the final questionnaire and during the debrief session provide insights on the benefits and weaknesses of the different enhanced-audio configurations. For example, some participants indicated that in the audio-only condition, the audio was disembodied and difficult to follow: "*The speeded up audio for me was difficult to understand sometimes*;" "*When it was audio only, I didn't have a place to look or focus at so I lost focus in the conversation. It was really difficult for me to focus on the audio*." One participant stated that anything in addition to audio would help because there were no clues to understand audio when it went too fast, "*There was no other reference. A lot of things kind of blew by me … I would have liked to have any other clue at that point*." In the enhanced-audio conditions, audio could be correlated with other information and hence it was easier to understand and remember what was said.

For the audio+video review, participants indicated that the video review made it easier to remember which presenter answered each question, "*Being able to see who was talking during catch-up helped to associate a face, name, and voice with the answers given,*" and "*Video kept my attention and enabled me to focus on what and who said it*." Others, however, found the video less useful, "*Video does not convey a whole lot*," and "*The video on the catch-up, that added nothing to the ability to pick up (information)*."

For the audio+workspace and audio+transcript reviews, participants used the shared workspace and transcript as a reference to what they were hearing. One participant in the audio+transcript condition found that she had trouble following the accents of the actors, and she used the transcript to double-check what she heard, "*The written language was essential because the verbal was not enough for me*." Others said that they were visual learners, "*I'm a visual learner, so I heard and remembered more*."

Some participants who used the audio+all review mechanism noted that they mainly used the shared workspace to cross reference what they heard. However, when that cross-referencing was not sufficient for full understanding of the audio, they also cross-referenced the audio with the transcript. As one participant put it, "*I used the presentation to answer the questions, I used the audio to see who was talking, and every time I missed something, I had the transcript which kept a recording of everything and I could just look back at it.*" Others simply used the transcript to cross reference the audio and ignored the shared workspace, "*It's easy to listen and watch with the ability to check the transcript for the things you miss*."

Overall, some participants simply said more is better, "*given the choice, the more information the better. It seemed very clear when you had all three of them (video, desktop, and workspace)*," while some others said that a minimalist approach works best "*I found that the more information that was there, the more confused I became. I think I did better during catch-up when I had audio-only*."

Transcript-Only

Of the ten participants that took part in the transcript-only condition, only two preferred transcript-only. Feedback from the participants indicated that they found it more difficult to keep up with the text during review in the transcript-only condition as compared to the accelerated audio in the audio-only condition.

Table 5 shows the recall results for the transcript-only condition compared to audio-only. A significant interaction effect of condition and question type was found ($F_{2,18}$=9.67, p<.001). Examining each question type, recall was significantly higher for facts in the transcript-only condition compared to audio-only ($F_{1,9}$=5.16, p=.049) but no significant difference was found for explanations ($F_{1,9}$=0.16, p=.70). Identification recall for the transcript-only condition was significantly worse than the audio-only condition ($F_{1,9}$=6.02, p=.037). Finally, there were no significant differences in the participants' ratings of their confidence (p>.05).

Recall	Fact	Explanation	Identification
Audio Only	71%	50%	56%
Transcript Only	77%[1]	52%	43%[2]

Table 5. Percentage of correct answers. [12]Transcript only was significantly better than audio only for facts, but was significantly worse than audio only for identification, p<.05.

Feedback from the participants indicated that they found it more difficult to keep up with the text during review in the transcript-only condition than the accelerated-audio in the audio-only condition. As they put it, "*I had difficulty following the scrolling*," "*I would much prefer to hear you talking faster. It was much easier to catch up than following the transcript*," and "*Transcript alone was way too fast to really understand what was going on. I could kind of skim it and I got some information but I felt like I was just bouncing along.*" Others realized later that they were not actually reading all of the information in the transcript, "*I would read the words, but I would forgot to read the names*." Several of them commented that it would have been better to allow manual scrolling through the transcript. While our mocked up version of AIR used in the study did not support this functionality, the real AIR system does.

Results Summary

The main results of this study demonstrate that enhanced-audio catch-up is superior to audio-only review. Both the subjective and objective results are consistent with this finding: overall, users preferred, felt more confident with, and performed better with enhanced-audio than with audio-only review. Closer examination of the individual modalities in the enhanced-audio conditions shows that compared to audio-only review, using all of the modalities to catch up showed the strongest benefit, with significantly higher user confidence and recall of facts, explanations, and identification. Additionally, using audio along with the shared workspace and transcript modalities for catching up significantly improved the recall of facts and explanations, respectively, compared to using audio-only. Finally, results from the transcript-only condition showed that this condition is slightly better than audio-only for fact recall, but was significantly worse for speaker identification. In terms of preference however, most participants did not like the transcript-only condition.

DISCUSSION

We have presented results that show the benefit of the AIR Conferencing system in a practical scenario – a status update meeting – when user absences are brief (a few minutes) and the number of users is small (three). In this section, we address our contributions with respect to their ecological validity and generalizability, and discuss other important issues concerning in-meeting review systems.

Data from the interviews in our first study indicate that a system like AIR would be useful in real meetings. Specifically, users reported that they arrive late for meetings or get interrupted, and they want to be able to catch up on information they missed. They also reported that these interruptions were relatively short (less than five minutes). In addition, most of the interviewed users had previous experience with both videoconferencing systems and DVR systems, and see benefits to both. Given that people frequently take advantage of DVR-like functionality (e.g., pause, replay, rewind) when they miss something, or do not understand something when watching television, it is plausible that this type of behavior could easily transition to videoconferencing. When asked about this possibility, most of the users in the study strongly felt that these features would be useful. Overall, these findings suggest that an accelerated review technique for in-meeting review could be useful for reviewing missed content.

Ecological Validity

In general, attaining ecological validity is a difficult task. The studies presented in this paper were designed to evaluate the usefulness of an in-meeting review system. Therefore, it was important to select a task that was both realistic and would allow for an objective evaluation. Given the conflicting nature of these requirements, we chose to focus on subparts of meetings that are common to many different types of meetings. Specifically, meetings typically have periods of presentation and recall. Our studies targeted these meeting subtasks. For instance, during presentations, facts are often presented, and when a participant wants to ask the presenter a question, the participant must recall some of these facts. During brainstorming sessions, factual data is often presented, such as the explanation of an idea or the reasons for pruning or choosing an idea. As brainstorming meetings are more interactive, people must recall not only facts and explanations, but also the participant who presented them.

Generalizability

The task we used in our studies focused on presentation and recall, which are common sub-tasks of real meetings; however, the structure of these in our task was simple and modular, consisting of alternating presentation and recall periods. While some real-world meetings are modular, many have a more complex structure. We did not evaluate how well in-meeting review systems support more complex meeting structures.

In addition to controlling the meeting structure, our study also controlled the interaction style. In particular, the non-presenting attendees were passive participants. They did not interrupt the presenter and were specifically told not to ask the presenter to catch them up when they returned from an interruption. In some real-world meetings, attendees are passive; however, in other kinds of meetings, all attendees have a stake in the outcome and are actively participating. In this case, an attendee may want to jump from replay to live if something interesting comes up in the live discussion. Similarly, a reviewing attendee may want to jump to the live meeting when asked a question. Handling these situations requires that users divide their attention between past and live content. In our first study, some participants did so by reading the transcript of one and listening to live audio of the other.

In general, participants could not act on divided attention cues because we controlled the replay to ensure consistency between participants and across conditions. Our participants always reviewed all content from the point in time at which they were interrupted and were not allowed to jump to the live discussion manually; instead, they had to wait for the replay mechanism to catch up to live. These replay restrictions allowed us to study replay effectiveness in absence of variables arising from control of the mechanisms; however, letting users choose when, what, and for how long to review would provide insights on how users multitask in meetings and help us better understand issues of divided attention.

Other Questions

The results of this work demonstrate the potential of in-meeting review; however, there are several questions that have not been answered by this work. One question is the social impact of review systems. For instance, in some cases, such as a fully-distributed videoconference, the review can be done privately. In such cases, in-meeting review may not have any social impact. However, in other cases, such as face-to-face meetings, the fact that an attendee is reviewing is going to be obvious to other attendees, which may have negative side effects. For instance, others may think that the reviewing attendee feels the current discussion is not important and is choosing not to participate live. In-meeting review systems may also impact meeting dynamics. For instance, what will happen if multiple attendees are reviewing at the same time? How is the conversation impacted? At what point does the meeting break down? What happens if all participants are reviewing? Before in-meeting review systems are introduced in real-world settings, the impact on social and meeting dynamics needs to be evaluated further.

Finally, we studied several modalities and their combinations as catch-up mechanisms. However, we did not study all possible combinations of modalities. Of particular interest is the combination that reviews everything but video. Since video is typically the most expensive modality in terms of computation and bandwidth resources, if lack of video does not significantly alter user preferences and performance, then future systems could review video only when resources are abundant. Also, some of the catch-up mechanisms we studied provided users with a computer generated speech-to-text transcript. Unfortunately, current state of the art speech-to-text systems require several hours of training on a per user basis

to generate a fairly accurate (but not necessarily perfect) transcript, which may be a barrier for the adoption of speech-to-text systems. Until speech-to-text systems improve so that they are accurate with little or no training, the use of speech-to-text transcripts for catch-up purposes is going to be limited. Eventually, when they do improve, the use of audio+transcript with or without any additional modalities may become a viable catch-up mechanism.

CONCLUDING REMARKS AND FUTURE WORK

This work makes several significant contributions. We show that users prefer reviewing audio along with any additional modality to reviewing audio alone. We also show they are most confident and perform best when audio is reviewed simultaneously with video, shared workspace, and conversation transcript. Additionally, they had better recall of facts and explanations when reviewing audio along with the shared workspace and transcript, respectively, compared to reviewing audio only. However, when reviewing video along with audio, they performed similarly to when reviewing audio only. Also, the transcript-only review improved their recall of facts but degraded their recall of speaker identification compared to audio-only review. Finally, we designed a new in-meeting review system that goes beyond replaying just audio by incorporating audio, video, shared workspace actions, and a speech-to-text transcript into an accelerated review and demonstrated its usefulness in a live videoconference.

Our work also suggests several design considerations for future catch-up systems. 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. If either of these two dimensions is off, users may not utilize the system. Additionally, users are concerned about disrupting the meeting or missing more of the meeting when trying to catch-up. Care must be taken when designing the user experience to ensure that the process is seamless and does not detract from or disrupt the flow of the meeting.

We have explored several catch-up mechanisms in our AIR system; however, we have barely scratched the surface of in-meeting support. We plan to explore the potential of catch-up mechanisms that use different combinations of the modalities we studied and other new modalities, such as spatial audio and shared workspaces that support pen and touch interaction. We also plan to study how well users can attend to both the past and the present through audio spatialization by leveraging the cocktail party effect. In addition, we intend to evaluate the social impact of in-meting review and study how it affects meeting dynamics.

REFERENCES

1. Christel, M. Evaluation and user studies with respect to video summarization and browsing. *Symposium on Eletronic Imaging 2006*, 17-19.
2. Cutler, R., Rui, Y., Gupta, A., Cadiz, JJ., Tashev, I., He, L. Colburn, A. Zhang, Z., Liu, Z. Silverberg, S. Distributed meetings: A meeting capture and broadcasting system. *ACM Multimedia 2002*, 503-512.
3. Dietz., P. H., and Yerazunis, W.S. Real-time audio buffering for telephone applications. *ACM UIST 2001*, 93-94.
4. He, L., Sanocki, E., Gupta, A., and Grudin, J. Auto-summarization of audio-video presentations. *ACM Multimedia 1999*, 489-498.
5. Inkpen, K., Hegde, R., Junuzovic, S., Brooks, C., Tang, J., and Zhang, Z. AIR Conferencing: Accelerated instant replay for in-meeting multimodal review. *ACM Multimedia 2010*. 663-666.
6. Kazman, R., Al-Halimi, R., Hunt, W., and Mantei, M. Four Paradigms for Indexing Video Conferences. *IEEE Multimedia 3,1 (1996),* 63-73.
7. Meetings in America V: meeting of the minds. https://e-meetings.verizonbusiness.com/global/en/meetingsinamerica/uswhitepaper.php (2003).
8. Money, A. G., and Agius, H. Video summarization: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Recognition 19 (2008)*, 121-143.
9. Moran, T., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., van Melle, W., and Zellweger, P. I'll get that off the audio: A case study of salvaging multimedia meeting records. *ACM CHI 1997*, 202-209.
10. Omoigui, N., He, L., Gupta, A., Grudin, J., and Sanocki, E. Time-compression: Systems concerns, usage, and benefits. *ACM CHI 1999*, 136-143.
11. Orr, D.B. A perspective on the perception of time compressed speech. In P.M. Kjeldergaard, D.L. Horton, & J.J. Jenkins, (Eds.) Perception of Language, 108-119, Merrill, 1971.
12. Ranjan, A., Birnholtz, J. and Balakrishnan, R. Improving Meeting Capture by Applying Television Production Principles with Audio and Motion Detection. *ACM CHI 2008*, 227-236.
13. Tucker, S., and Whittaker, S. Accessing Multimodal Meeting Data: Systems, Problems, and Possibilities. Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 2004.
14. Tucker, S., Bergman, O., Ramoorthy, A., and Whittaker, S. Catchup: A useful application of time-travel in meetings. *ACM CSCW 2010*.
15. Wildemuth, B., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., Gruss, R. How fast is too fast? Evaluating fast forward surrogates for digital video. *ACM/IEEE-CS Digital Libraries 2003*, 221-230.