Mediated Meeting Interaction for Teleconferencing

Kazumasa Murai, Don Kimber, Jon Foote, Qiong Liu, John Doherty
FXPAL Japan, FXPAL
kazumasa.murai@fujixerox.co.jp, {kimber, foote, liu, doherty}@fxpal.com

Abstract

A common problem with teleconferences is awkward turn-taking – particularly 'collisions,' whereby multiple parties inadvertently speak over each other due to communication delays. We propose a model for teleconference discussions including the effects of delays, and describe tools that can improve the quality of those interactions. We describe an interface to gently provide latency awareness, and to give advanced notice of 'incoming speech' to help participants avoid collisions. This is possible when codec latencies are significant, or when a low bandwidth side channel or out-of-band signaling is available with lower latency than the primary video channel. We report on results of simulations, and of experiments carried out with transpacific meetings, that demonstrate these tools can improve the quality of teleconference discussions.

1. Introduction

A common problem in teleconferences is awkward turn-taking. Time is wasted as participants collide with each other while initiating turns at speaking, or wait for each other to start. More importantly, the flow of the discussion is interfered with, resulting in a degradation of the quality of interaction compared with face-to-face meetings. Multiple factors contribute to the problem, including latency, degraded audio video quality compared with face-to-face meetings, and cultural differences. In this work, we focus on problems resulting from latency in the transmitted audio/video signal due to coding/decoding and network transmission delays. In transpacific meetings between our research centers in the US and Japan, using an MPEG2 based teleconferencing system, for example, the delay in each direction is about 700msec. While high-end state-of-the-art teleconferencing systems provide lower encoding/decoding latency, network delays are unavoidable.

We use simple models to analyze the effect of latency on collisions, and propose visual interfaces for helping meeting participants in turn-taking, particularly in avoiding collisions. One such interface is shown in Figure 1, and assists users in awareness of latency by visually indicating when the 'channel has cleared'. The graphic at the bottom right hand side is a series of audio level bars, with successively greater time delays, depicting the local audio 'traveling through cyberspace' to remote sites. The final bar is delayed by the total latency time to remote sites and back. After a local turn ends, the time at which that bar clears is the earliest time that a non-interrupting reply could be expected from the remote site. There is also a traveling wave depicting an incoming signal, which can be seen before the remote audio signal is heard, because it is based on a lower latency side channel, which need not be delayed for synchronization with video.

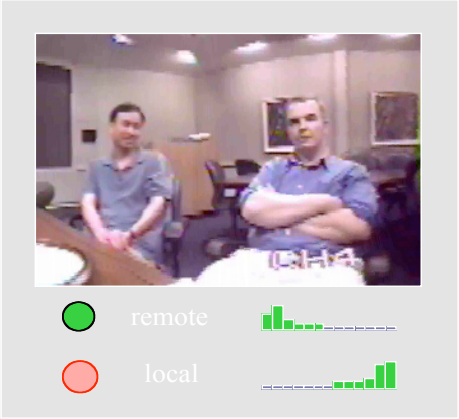

Figure 1. Augmented Interface showing speaker state and channel state.

1.1 Related Work

There is a rich literature on conversation analysis, involving sociological studies of turn-taking behavior. [1] Ruhleder and Jordan review some of that work, and provide qualitative microanalysis of video based interaction and discuss delay related problems. [2] Sellen compares speech patterns in face-to-face conversations with those in low latency video-mediated conversations. [3] Fischer and Tenbrink discuss turn-taking in video-conferences, and describe floor control strategies adopted by meeting participants. [4] Those studies are descriptive however, and do not propose statistical models for conversational properties such as

0-7803-9332-5/05/$20.00 ©2005 IEEE

collision rates, nor do they suggest interface enhancements to improve conversation characteristics.

Another related area is in floor control in collaborative systems. Katrinis et al. review several such systems and propose a floor control protocol, but focus on concurrency control for shared multimedia resources. [5] Other authors have worked on systems with explicit management of turns, such as through token passing. [6][7] Our system differs by assisting with turn-taking through enhanced awareness rather than explicitly imposing formal floor control.

2. Analysis and model

Many aspects of teleconference meeting dynamics can be analyzed in terms of state transition diagrams, as seen in Figure 2, which illustrates a discussion between two speakers; Ann at Site A, and Bob at site B. In simplest form, the state of a participant is characterized as Silent or Active (i.e. talking.) In Figure 2, Ann first begins speaking, and a time ΔT_1 later, she is heard at Site B. After she finishes, Bob replies, and time ΔT_2 later, Ann hears his reply. Note that when a speaker completes a turn, the earliest they should expect to start hearing a reply is after $\Delta T = \Delta T_1 + \Delta T_2$. After Bob's turn, Ann immediately replies. However, before Bob hears Ann, he begins talking again, resulting in a collision. This collision may have been avoided if Bob had been more aware of the channel latency.

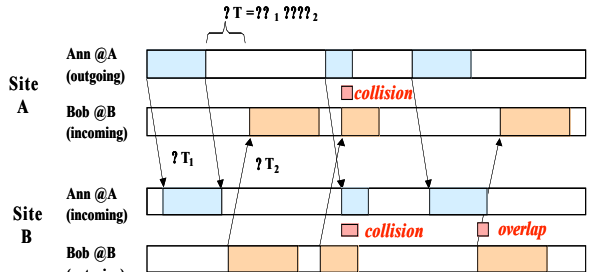

Figure 2. State-time diagram of teleconferenced discussion.

We define a collision as an event for which a speaker at a site is active at the same time the incoming channel is active, (i.e. that a remote participant is heard.) This condition can occur in several ways. One is that after a period of silence, multiple speakers begin talking before they become aware someone else is talking. Typically when such a collision occurs, after a short time both speakers stop talking. Another case is that someone misjudges the end of another person's turn, and begins talking before they had really finished. Another case is when a participant intentionally interrupts a speaker. Usually when a collision occurs at one site, it also occurs at remote sites, but this is not always true. The later part of Figure 2 shows a situation in which a collision occurs at B, but not at A.

2.1 Stochastic model

Our simulation model is illustrated in Figure 3. The state of participants is characterized as Active or Silent. The 'local state' at a site is characterized as Active or Silent, according to whether a local participant is talking. The 'channel state' at a site is characterized as Active or Silent, according to whether a remote participant's turn is being played. The state of a site is characterized by the local participant state at the site, indicated by a small circle in Figure 3, and the incoming channel state, indicated by a small box. In our model, the state at a site changes for one of two reasons. (1) A person at that site changes state, (2) the channel state at the site changes. The state of the overall teleconference is characterized by the states of all sites, and the state of the channels, including the time delays between sites.

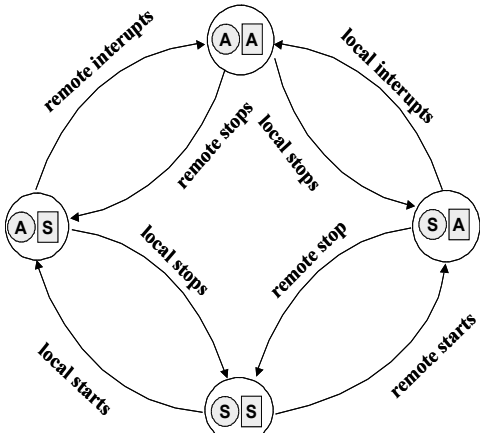

Figure 3. Simplified State diagram for two speakers.

A model of speaker behavior could be arbitrarily complex and take into account cognitive state, media content, etc. However, even highly simplified speaker models provide insight into collision characteristics, such as how collision rates increase with delay time or with increased numbers of sites. We have used a very simple model where after a state change at a site, a speaker will wait a randomly distributed amount of time before spontaneously changing state. For example, if a site enters the state (speaker=Silent, channel=Silent) we model a speaker as waiting a random time T_{SS} before beginning to talk. Before that time elapses, the state may change due to an incoming channel signal, so the state becomes (speaker=Silent, channel=Active). Then the speaker would wait a new random amount of time T_{SA} before they would

interrupt. Each of these random variables is described by some probability distribution. The cases are:

T_{SS}	Time to wait during silent period before talking.
T_{AS}	Time to keep talking while the channel remains silent.
T_{AA}	Time to keep talking after the channel becomes active.
T_{SA}	Time to listen to remote speakers before interrupting.

To understand collision behavior as a function of channel delay and number of sites, we ran simulations based on this model. For a simplified case, we consider T_{AA}=0, meaning that when a speaker is interrupted, they immediately yield, and T_{SA} =Infinity, meaning a speaker is 'polite' and never interrupts. For T_{SS} and T_{AS} we used an exponential distribution with mean values μ_{SS} =4 and μ_{AS} =20. The rates of collisions for those simulations are shown in Figure 4 for delays from 0 to 3 seconds, and for numbers of sites N, from 2 to 6. Simulations carried out with other distributions, such as uniform, resulted in qualitatively similar graphs.

Figure 4. Collision Data for Model

In some situations the probability of collision can be described in closed form. For example after a long silence, multiple speakers may begin turns at times within the channel delay. For exponential random variables T_{SS} with mean μ_{SS} the probability of this is:

$$\Pr(\text{Col}) = 1 - (1 - F_{SS}(\Delta T))^{N-1}$$
$$= 1 - (1 - \exp[-\Delta T / \mu_{SS}])^{N-1}$$

where $F_{SS}(t) = \Pr\{T_{SS} < t\}$ is the CDF, $(1-\exp[-t/\mu_{SS}])$.

3. Interface and implementation

We have implemented a 'traveling wave' audio level meter, which is used in conjunction with the existing teleconferencing system. The system uses a single video camera and an echo-canceling microphone and speaker at each end, and MPEG2 based hardware codecs. Synchronized audio and video is delivered over the network with a latency of 700mS in each direction.

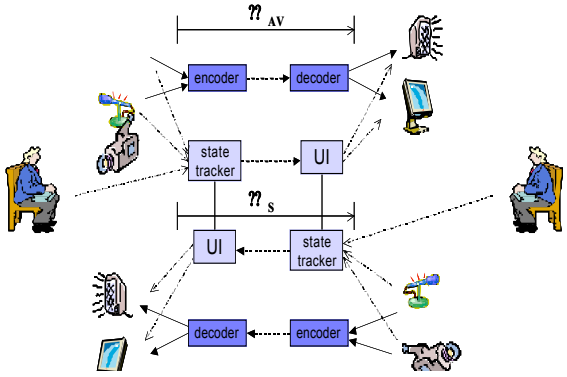

Figure 5. System Architecture.

For our enhanced interface, we split the audio signal from the echo-canceling microphone output at each site, send an averaged signal level to the other site over the same TCP/IP network, and display the audio signal level at the receiver's teleconferencing video device. Because the audio level information does not require compression, and is not delayed for synchronization with the video, the level meter is displayed with about 240mS less latency than the actual audio signal, i.e. the level meter precedes 240mS ahead the speech.

Figure 1 shows a visual indicator for the state at each site, Active or Silent, and a 'traveling wave' showing the levels changing with time. The level shown for the local signal does not require any transmission, and simply requires the system to know the delay. The remote level is seen as an 'approaching wave' that reaches the current site at just the time the audio is heard. In future systems, additional information may be shown, such as the cumulative duration of speech by various participants, or an indication that a speaker has requested the floor.

4. Experiments

We conducted experiments to verify the effectiveness of our interface, by comparing collision properties with and without the proposed system attached to an existing teleconferencing setup between sites in the US and Japan. The audio and video latency is 1,400mS round trip, including CODEC and

TCP/IP latency, while the TCP/IP latency alone is about 120mS round trip.

4.1 Experimental conditions

As a control setup, 3 pairs of Japanese native speakers were instructed to discuss, *without the proposed system*, the arrangement of a weekend tour (or a welcome party) under the assumption that one examinee is making a business trip to the other participant's office. A few days later, *using the proposed system*, the same pairs are instructed to discuss the same topic, but the travel destination is the other way. Before the latter session, the examiner briefly described the level meter to the examinees. For evaluation purposes, we used speech information recorded with a close-talk microphone.

4.2 Evaluation results

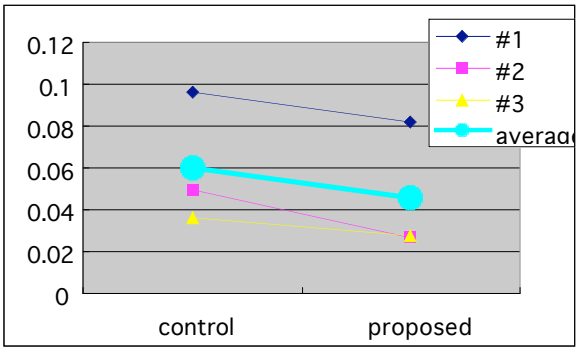

Figure 6. Collision counts per second

After cleaning up the audio channel by excluding the microphone noise and other noise manually, each utterance was hand-labeled. By analyzing the audio stream at each side, collisions were labeled, and collision rates determined. Figure 6 shows the rates of collisions for the three pairs of speakers, and the average rate, with and without the proposed interface, based on the labeled data for the Japan side. The rate of collisions decreased in each case.

4.3 Usability

We have interviewed all examinees about the usability of the proposed system and found that:

1. The system does not require a user to keep watching the meter, although the user may use it whenever he/she is willing to see it.
2. The 'traveling wave' level meter should be improved so that a level reaches the end of the wave exactly when the corresponding audio is played. In the experiment, the wave was too long and showed levels for audio that had already been played.

In addition to these comments, all of the examinees are willing to use the system for their own teleconferencing.

We also used our tool in weekly meetings. One surprise was that the interface seems especially useful for avoiding collisions resulting from misjudging the ends of remote speaker turns. After hearing a pause in the audio, by viewing the interface, it is possible to know if the remote speaker is about to continue.

5. Conclusions

We have proposed and evaluated a teleconferencing system based on an analysis of collisions. Our system provides visual cues to help mediate turn-taking and reduce collisions. Experiment results showed that (1) the collision rate is higher than our simulation result, (2) the proposed system assists participants by reducing the rate of speech collisions, and (3) all participants are favorable to using the system. The visual cues may have assisted participants with awareness of latency, and/or to noticing the remote participant being about to initiate a turn.

6. References

[1] H. Sacks, E. A. Schegloff and G. Jefferson: "A Simplest Systematic for the Organization of Turn-Taking for Conversation," *Language*, 50, 4 (1974) 696-735 .

[2] K. Ruhleder and B. Jordan, "Co-constructing non-mutual realities: delay-generated trouble in distributed interaction." *Computer-Supported Cooperative Work*. 2001; 10 (1): 113-138.

[3] Abigail Sellen, "Speech patterns in video-mediated conversations", *Proceedings of CHI '92*, ACM, 1992, pp. 49-59.

[4]] K. Fischer and T. Tenbrink, "Video conferencing in a transregional research cooperation: Turn-taking in a new medium" http://nats-www.informatik.uni-hamburg.de/~fischer/VKfischertenbrink.pdf

[5] K. Katrinis, G. Parissidis and B. Plattner, "Activity Sensing Floor Control in Multimedia Collaborative Applications," http://citeseer.ist.psu.ed/716475.html

[6] A. Patrick, "The Human Factors of MBone Videoconferences: Recommendations for Improving Sessions and Software", *Journal of Computer-Mediated Communication,* 4(3), 1999.

[7] H.P. Dommel and J.J. Aceves, "Floor Control for Multimedia Conferencing and Collaboration", *ACM Multimedia Systems*, Vol. 5, No. 1, 1997, pp. 23-38.