Video Increases the Perception of Naturalness During Remote Interactions with Latency

Jennifer Tam
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
jdtam@cs.cmu.edu

Elizabeth J. Carter
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
ejcarter@andrew.cmu.edu

Sara Kiesler
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
kiesler@cs.cmu.edu

Jessica K. Hodgins
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
jkh@cs.cmu.edu

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
Visual telecommunication systems support natural interaction by allowing users to remotely interact with one another using natural speech and movement. Network connections and computation cause delays that may result in interactions that feel unnatural or belabored. In an experiment using an audiovisual telecommunications device, synchronized audio and video delays were added to participants' conversations to determine how delay would affect conversation. To examine the effects of visual information on conversation, we also compared the audiovisual trials to trials in which participants were presented only the audio information. We present self-report data indicating that delay had a weaker impact when both audio and video channels were available, for delays up to 500 ms, than when only the audio channel was available.

Keywords
delay; latency; telecommunication; audio; video

ACM Classification Keywords
H.5.1 [Multimedia information systems]: Evaluation;

Introduction
Visual telecommunication systems are popular because they support more natural forms of interaction than

Authors	Delay (ms)	Pair Types
Riesz and Klemmer, 1963 [10]	600	co-workers
Klemmer, 1967 [5]	600	co-workers
Krauss and Bricker, 1967 [6]	900	strangers
Kitawaki and Itoh, 1991 [4]	560	co-workers and strangers
Kurita, et al., 1994 [7]	300	co-workers
Holub, et al., 2007 [2]	500	strangers

Table 1: Audio delay thresholds found in prior work.

telephones and text-based chat rooms. Non-verbal behaviors, such as head nods, facial expressions, eye blinks, eye gaze, and lip movements, are available during interaction, and these behaviors are important cues that improve the ability to express understanding, agreement, and attitude, enhance verbal descriptions, interpret pauses, and take turns [1, 3]. Unfortunately, visual telecommunication systems are subject to network and computational delays, which can negatively impact users' interactions. Audio delays cause people to interrupt each other more frequently and to spend more time gaining control of or clarifying the conversation [7, 11]. Industry experts suggest that audio delays be below 200 ms [8, 9]. Many researchers have identified thresholds at which delay becomes noticeable or interferes with audial conversation, but these thresholds are inconsistent (see Table 1). Although all delay thresholds were determined based on free conversation tasks, the thresholds range from 300-900 ms. The studies conducted in English [5, 6, 10] suggest that the delay threshold is within 600-900 ms, but these are also the oldest studies. The newer studies [2, 4, 7], which also happen to be in non-English languages, suggest that the delay threshold is within 300-560 ms. Besides the different language, these lower thresholds may be due to the fact that participants were directly asked about delay and interference.

Prior research investigated the differences between audio and audiovisual platforms in regards to communicative efficiency and noticeability of delay. Isaacs and Tang [3] evaluated the differences in interaction between individuals collaborating on a task using both audio and audiovisual telecommunications systems . They found that the addition of video allowed participants to better understand each other and express themselves. Turn taking within the conversations was easier with video than without, and overall, the interactions were considered easier than the audio only interactions. Kurita and colleages [7] examined the noticeability of delay with participants who used both audio and audiovisual systems. They found that there were no differences between participants' perceptions of delay regardless of which system they used. We investigated the differences between audio and audiovisual platforms in terms of the quality of interaction.

Hypothesis

Prior research suggested that delays would be noticeable between 200 and 600 ms. To discover exactly how much delay it would take to negatively impact conversations, each participant was exposed to seven different delay conditions between 67 and 900 ms. We expected that long delays would cause conversations to feel unnatural and uncomfortable.

Because prior research had not reached a clear conclusion regarding the possible benefits of visual information in the presence of delay, we also investigated the differences in conversational attribute ratings between audio and audiovisual conversations. We expected that participants using the audiovisual system would experience a more natural and more comfortable conversation than those using the audio system because nonverbal information is so important in normal conversation. We hypothesized that delay would have less negative effect on the naturalness and comfortableness of the conversation if video were available. Prior research also suggested that audio delay increased the number of interruptions in a conversation, but because non-verbal information is so important to turn taking [1, 3], we expected that delay would not have as much of an effect on the number of perceived interruptions when participants could see one another.

Trial	Topic
1	favorite food
2	favorite vacation
3	hobby
4	dream vacation home
5	event to plan
6	favorite restaurant
7	activity to try

Table 2: Topics used in study.

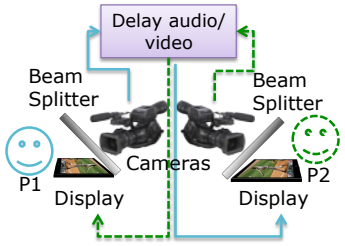

Figure 1: Audiovisual telecommunications device setup.

Materials and Method

We examined conversational attribute ratings in a controlled laboratory experiment for adult, native English speakers. Each pair of participants had seven conversations about selected topics, each of which were followed by short surveys asking participants to rate the conversation on various attributes. Half of the pairs conversed with both *audio and video*, while the other half conversed with *audio only*. Different amounts of delay were inserted into each of the seven conversations.

Participants

We advertised our study on a university experiment scheduling site. Fifty-six adults participated in this study (age range: 18-59 years; median age: 24 years; 28 females). Participants, who were strangers to one another, were run in same-gender pairs. All participants completed informed consent forms approved by the Institutional Review Board. Participants were paid $15 for the hour long study.

Apparatus

Our audiovisual telecommunications system consisted of two stations located in separate rooms (see Figure 1). Each station consisted of a 12-in.$\times$14-in. beam splitter contained in a black box. The beam splitter allowed participants to make eye contact without seeing the camera. Audio Technica shotgun microphones recorded audio, and audio delays were controlled via a Yamaha 01v96 digital audio mixer. Video was captured by an AJA Kona card in an Apple 6-core 2$\times$2.93 GHz MacPro running software that could delay sending the frames to the monitor of the beam splitter. The intrinsic delay between the two stations was 67 ms. Calibration of the video delay software ensured audio and video synchronization were maintained throughout the study.

Experimental design

We used a repeated measures experimental design. Delay was a within-subjects factor with seven conditions (67, 200, 300, 400, 500, 600, 900 ms), and communication channel (CC) was a between-subjects factor with two conditions (*audio only* or *audio and video*). Delay conditions were chosen based on previous research and extensive pilot testing. In the *audio only* CC condition, participants saw a static desktop image of a purple sky on the screen. The delay conditions were assigned to participant pairs in a 7×7 Latin square design. Participants were randomly assigned to a CC condition. Finally, based on their CC condition, a pair was then assigned to one of four Latin squares, resulting in one square for each CC condition and gender combination.

The topic ordering was kept the same across all participant pairs (see Table 2). Before each conversation, participants were given topic sheets that included some sample questions and basic prompts that could be used to keep the conversation alive.

Procedure

Each pair of participants completed consent forms at the study location. They were then taken to separate study rooms containing the audiovisual telecommunications stations. The experimenters informed the participants that they would have seven 4-minute conversations using the stations, and that they would be given topic sheets for inspiration. Participants could use a small timer to keep track of their conversation, and they were informed that the experimenter would interrupt the conversation once four minutes had passed. Once seated, partcipants were given headphones and the first topic sheet. Participants were told to start whenever they both were ready. The experimenters then left the study rooms to monitor the conversations from a nearby location.

Scale	Questionnaire Items	Alpha*
Topic likeability	Did you like or dislike the topic? Do you think your partner liked or disliked the topic?	0.8646
Comfortableness	How comfortable or uncomfortable did you feel? Did you find your partner comfortable or uncomfortable?	0.8875
Naturalness	How was the flow of this conversation? How natural or unnatural did you find this conversation? Was this conversation like or unlike an in-person conversation	0.8585
Perceived interruptions	How many times did you and your partner interrupt one another?	NA
Perceived pace	How quick or slow was your partner to respond?	NA

Table 3: Questionnaire items administered after each conversation. *Cronbach's Alpha is a measure of the reliability of the scale as a whole. Alpha ranges from 0.0 to 1.0.

After four minutes, the experimenters interrupted the conversations, gave the participants short surveys to complete, and presented the next topic sheet. This process was repeated for each of the seven trials. After all seven trials, participants completed a questionnaire asking about their favorite conversations and any difficulties with understanding the other participant.

Measures

Immediately following each conversation, participants rated the conversation on nine, five-point scale items. We chose the questions to reflect our main interest in the flow of conversation and how delay might disrupt conversation. We combined some items after exploratory factor analysis suggested they loaded on the same factor. Table 3 lists the questions.

Results

The self-report data indicated that video weakened the negative impact of delays on naturalness for delays up to 500 ms, whereas in conversations with no video, delays at or above 400 ms negatively impacted naturalness. Once delays were at or above 600 ms, conversations from both CC conditions were perceived as significantly less likeable, comfortable, and natural. Interruptions increased with delay and were not mitigated by the addition of video.

Effects of delay

We found delay to have a significant impact on all scale items except for pace which was only marginally affected (see Figure 2). We discovered that as delays increased, likeability of topic decreased, $F(6, 318) = 2.43$, $p = .03$. Participants especially disliked topics presented with delays at of above 600 ms compared to those presented with shorter delays, $F(1, 318) = 6.15$, $p = .01$. Delay also had a significant effect on comfortableness, $F(6, 318) = 2.29$, $p = .04$, and naturalness , $F(6, 318) = 3.29$, $p = .004$, with both qualities decreasing with the increase of delay. We expected long delays would cause conversations to feel unnatural and uncomfortable. When conversations were presented with delays at or above 600 ms, they were rated significantly more unnatural, $F(1, 318) = 16.95$, $p < .0001$, and uncomfortable, $F(1, 318) = 6.95$, $p = .009$, than conversations with delays between 67-500 ms.

Interruptions significantly increased with delay, $F(6, 318) = 6.56$, $p < .0001$, and as depicted in Figure 2, the amount of delay was found to be significantly correlated to the number of interruptions ($r(392) = 0.1891$, $p < .001$). Overall, conversation pace was only marginally affected by delay, $F(6, 318) = 1.64$, $p = .14$, but when delays were at or above 600 ms, the

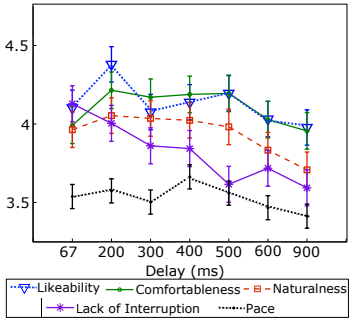

Figure 2: Effects of delay on conversation perceptions. The main effect of delay on perceptions across all scales except for pace is statistically significant, $p < .05$.

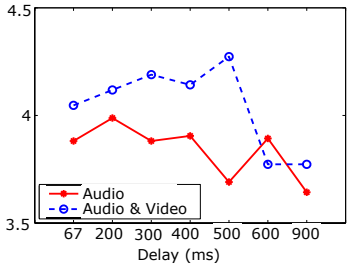

Figure 3: Effects of delay and channel on perceived conversation naturalness. The interaction effect of delay and channel on the perception of naturalness is statistically significant, $p < .05$.

pace of the conversation was considered to be significantly slower than conversations with delays between 67-500 ms, $F(1, 318) = 5.82$, $p = .02$.

We found an interaction effect of communication channel and delay on naturalness ($F(6, 318) = 2.27$, $p = .04$). Contrast tests revealed that conversations with short delays of 67-300 ms did not significantly differ in naturalness between the two CC conditions. Conversations with mid-length delays of 400 and 500 ms maintained their naturalness in the *audio and video* CC condition and decreased in naturalness in the *audio only* CC condition ($F(1, 77) = 3.84$, $p = .05$). Conversations with delays at or above 600 ms were the most unnatural ($F(1, 318) = 16.95$, $p < .0001$). In other words, video weakened the negative impact of delay on conversation naturalness for delays up to 500 ms while audio only conversations were negatively impacted with delays at or above 400 ms.

Prior research indicated that delay would negatively impact conversation or become noticeable at some point between 300 and 900 ms. We consistently found that across all of our conversation attributes, conversations with delays above 500 ms were negatively impacted. All participants were given the opportunities to comment on any technological or communicative difficulties during the study. Participants were also told the study's purpose after their conversations, and they were asked if they had noticed any delays. Only 16 of the 56 participants indicated that they were aware of any delay (28.6%), suggesting that most strangers conversing with one another will not notice delays above 500 ms. This difference from previous studies may indicate that people today are more accustomed to delay due to the popularity and widespread use of internet telephony and video chat.

Additional analyses

The topic of conversation had a significant main effect on topic likeability, $F(6, 318) = 11.69$, $p < 0.0001$, with "favorite food", "event to plan" and "dream vacation home" being the least favorite topics. We believed that participants might require the first trial, "favorite food," to become acquainted with the equipment and each other. If this were true then the first trial should score significantly lower than the other trials, including the other trials with disliked topics, however, as this was not true we kept the first trial in the rest of our analysis.

Conclusion

Prior research suggested that audio delays between 300-900 ms would not only be noticeable, but that the delay would also negatively impact remote interactions. In our experiment, strangers conversing with one another indicated that delays negatively impacted likeability of conversation topic, comfortableness, naturalness, pace, and interruptions. In particular, delays at or above 600 ms had a significantly stronger impact than delays between 67-500 ms. Video was found to actually weaken the negative impact of delay on naturalness for delays up to 500 ms, whereas audio only conversation naturalness suffered from delays at or above 400 ms. This difference could be due to the fact that audiovisual interaction allows participants to see nonverbal information. These findings are promising for those working on audiovisual telecommunications systems, as they allow for a manageable 500 ms of latency due to computation and network delays. We are currently analyzing the behavioral data from this experiment. We expect delay to affect behavior, and we wish to characterize how delay will affect behavior. We plan to examine head nods, eye blinks, utterance and pause lengths and quantity, laughter, and interruptions.

Acknowledgements

This study was funded by NSF grant CCF-0811450. Jennifer Tam is supported by an NSF graduate research fellowship. We thank Iain Matthews, Peter Carr, Kenichi Kumatani, Brooke Kelly, Chase Smith, and Disney Research Pittsburgh staff and interns for their suggestions and help with this study.

References

[1] Duncan, S. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology 23* (1972), 283–292.

[2] Holub, J., Kastner, M., and Tomiska, O. Delay effect on conversational quality in telecommunication networks: Do we mind? In *Wireless Telecommunications Symposium* (2007), 1–4.

[3] Isaacs, E. A., and Tang, J. C. What video can and can't do for collaboration: a case study. In *ACM MM* (1993), 199–206.

[4] Kitawaki, N., and Itoh, K. Pure delay effects on speech quality in telecommunications. *IEEE Journal on Selected Areas in Communications 9*, 4 (1991), 586–593.

[5] Klemmer, E. Subjective evaluation of transmission delay in telephone conversations. *The Bell System Technical Journal 46*, 6 (1967), 1141–1147.

[6] Krauss, R. M., and Bricker, P. D. Effects of transmission delay and access delay on the efficiency of verbal communication. *Journal of the Acoustical Society of America 41*, 2 (1967), 286–292.

[7] Kurita, T., Iai, S., and Kitawaki, N. Effects of transmission delay in audiovisual communication. *Electronics and Communications in Japan (Part 1: Communications) 77*, 3 (1994), 63–74.

[8] Percy, A. Understanding latency in IP telephony. Tech. rep., Brooktrout Technology, 1999.

[9] Polycom. Supporting real-time traffic: Preparing your IP network for video conferencing. Tech. rep., Polycom, 2006.

[10] Riesz, R. R., and Klemmer, E. Subjective evaluation of delay and echo suppressors in telephone communications. *The Bell System Technical Journal 42*, 6 (1963), 2919–2941.

[11] Vartabedian, A. The effects of transmission delay in four-wire teleconferencing. *The Bell System Technical Journal 45*, 10 (1966), 1673–1688.