Visual Filler: Facilitating Smooth Turn-Taking in Video Conferencing with Transmission Delay

Hiroaki Kawashima
Grad. Sch. of Informatics,
Kyoto University.
Yoshida-Honmachi, Sakyo,
Kyoto, 6068501 JAPAN
kawashima@i.kyoto-u.ac.jp

Takeshi Nishikawa
Grad. Sch. of Informatics,
Kyoto University.
Yoshida-Honmachi, Sakyo,
Kyoto, 6068501 JAPAN
takeshi@vision.kuee.kyoto-u.ac.jp

Takashi Matsuyama
Grad. Sch. of Informatics,
Kyoto University.
Yoshida-Honmachi, Sakyo,
Kyoto, 6068501 JAPAN
tm@i.kyoto-u.ac.jp

CHI 2008, April 5 – April 10, 2008, Florence, Italy
ACM 1-xxxxxxxxxxxxxxxxxx.

Abstract

Turn-taking in a smooth conversation is supported by the anticipation of the floor handover timing among participants. However, it becomes difficult to maintain natural turn-taking in video conferencing with transmission delays because the utterances and movements of each participant are presented to the others with a time lag, which often leads to a collision of utterances. In order to facilitate smooth communication over a video-conferencing system, we propose a novel method, “Visual Filler,” that fills temporal gaps in turn-taking caused by the existence of delays. Visual Filler overlays an artificial visual stimulus that has a function similar to that of filler sounds on a screen with participant images. We have evaluated the effectiveness of a Visual Filler for reducing the unnaturalness of turn-taking on a simulated dyadic dialog situation with a delay.

Keywords

Video conferencing, transmission delay, filler, smooth turn-taking, computer-mediated communication.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; H4.3. Communications Applications: Computer conferencing, teleconferencing, and videoconferencing.

Introduction

Video conferencing enables us to convey much of the visual information that characterizes face-to-face interactive conversations to remote participants. However, face-to-face conversations are still preferred to video conferencing in some cases (e.g., business meetings) in spite of extensive travelling, which requires considerable time and costs, because some of the nonverbal cues that are important for facilitating smooth communication are not preserved in many of the conferencing systems.

The problems regarding spatial cues are well-focused in existing studies. In particular, the issue of eye contact attracts many researchers [1,6,11,16] because gaze information plays a major role in expressing intention and taking turns [9]. Other spatial cues such as the visual field, screen resolution, and the appearance size of participants are taken into account to realize immersive telepresence [7,8].

Although these spatial cues are not negligible, temporal cues such as the duration of response latencies are also crucial to preserve smooth face-to-face conversations [10,13]. The participants of video conferencing often complain about the temporal gap of back-channel responses delayed by the network transmission among conferencing systems even if the temporal delay is about half a second. In addition, it becomes difficult to maintain natural turn-taking because the utterance and motion of each participant is presented to the others with a time lag, which induces a temporal gap in estimating the appropriate timing of turn-taking [4]. As a result, participants often feel that the handover of the floor has failed and make another speech that leads to a collision of turns or utterances.

In order to divert the speaker's attention from unnatural temporal gaps, we propose the use of visual stimuli overlaid on a screen of video-conferencing systems. Since the function of the stimuli is similar to filler sounds, which fill redundant pauses (temporal gaps) during turn exchanges, we refer to both the method and the visual stimuli as “Visual Filler.”

Visual Filler

Before introducing the method of Visual Fillers, we describe the turn-taking over a video-conferencing system with a transmission delay. Modeling the turn-taking organization for a conversation is a research topic in itself [3,9,12]; however, we here introduce a minimum model of turn-taking that is simple enough to describe the effectiveness of Visual Fillers.

In the following paragraphs, we assume that floor is released or acquired subjectively by each participant in order to represent the temporal gap of turns or utterances during turn exchanges. Therefore, it may so happen that more than two participants have the floor at the same time or no one has the floor for a brief period of time.

Face-to-Face situations with no delay

Let us consider a face-to-face conversation with no delay. Figure 1 shows a single turn-taking between two speakers. First, the preceding speaker attempts to release the floor consciously or subconsciously by giving verbal, nonverbal, or paralinguistic cues that indicate the so-called transition-relevance places (TRPs) [12]. Then, the succeeding speaker recognizes the TRP and makes an utterance with a brief pause, which we refer to as a *transition interval*. We assume that the natural lengths of transition intervals are shared by the speakers and that these lengths depend on the context of conversations. For example, speech functions such as question, answer, agreement, and disagreement, which are often categorized as *dialog acts* [2,13], may influence these lengths. Finally, the preceding speaker recognizes the floor acquisition by the succeeding speaker, and there is a small temporal gap between the timing of the awareness and the timing pre-estimated by the preceding speaker. Thus, the handover of the floor is successfully completed, and the speakers' utterances rarely have long pauses or overlaps.

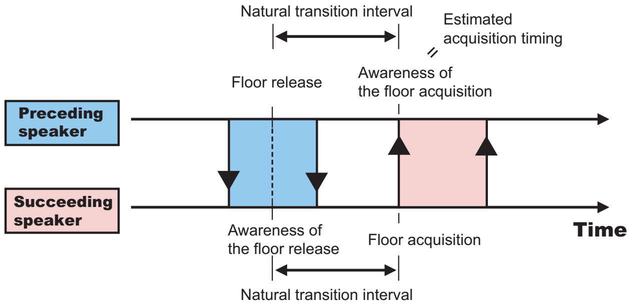
Natural transition interval
Estimated acquisition timing
Floor release
Awareness of the floor acquisition
Preceding speaker
Succeeding speaker
Awareness of the floor release
Floor acquisition
Time
Natural transition interval

figure 1. Turn-taking in a face-to-face conversation.

Video conferencing situations with delay
We then consider a video conferencing situation that has a certain length of transmission delay. The delay often increases the temporal gaps of turn exchanges to reach unacceptable levels. To simplify the model, let us assume a constant delay of length D. As shown in figure 2, the preceding speaker does not recognize the floor acquisition of the succeeding speaker at the timing he or she expected. The temporal gap in this case becomes $2D$. For example, let D be 0.5 s; then, the gap becomes 1 s, which is sometimes enough for the preceding speaker to feel that the succeeding speaker does not intend to hold the floor. As a result, the speaker makes another utterance. This leads to a collision of utterances; that is, the conversation will temporarily break down and be in need of repair.

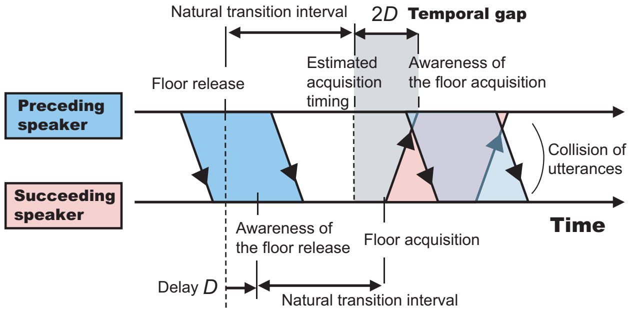
Natural transition interval
2D Temporal gap
Estimated acquisition timing
Awareness of the floor acquisition
Floor release
Preceding speaker
Succeeding speaker
Collision of utterances
Time
Awareness of the floor release
Floor acquisition
Delay D
Natural transition interval

figure 2. Turn-taking in video conferencing with delay D. The temporal gap from the natural transition interval becomes $2D$.

Visual Filler to fill temporal gaps
The idea of Visual Fillers is depicted in figure 3. To divert the preceding speaker's attention from unnatural temporal gaps, we suppose that the conferencing system inserts a Visual Filler, which is an artificial visual stimulus overlaid on the preceding speaker's screen with the images of the succeeding speaker.

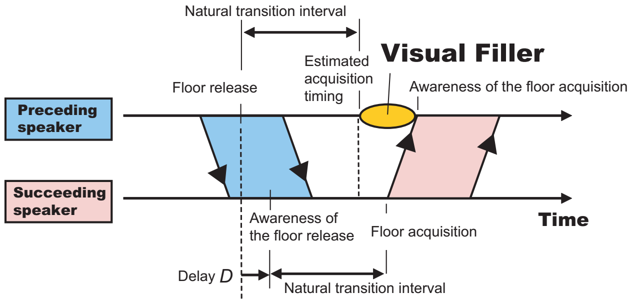
Natural transition interval
Visual Filler
Estimated acquisition timing
Floor release
Awareness of the floor acquisition
Preceding speaker
Succeeding speaker
Awareness of the floor release
Floor acquisition
Time
Delay D
Natural transition interval

figure 3. Insertion of the Visual Filler to fill the unnatural gap.

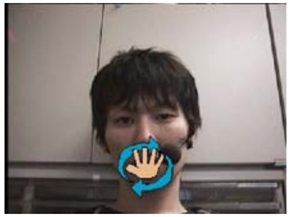

figure 4. An example of Visual Fillers used in the evaluation.

table 1. Eleven conditions used in the experiment. Each number denotes an interval length from the end of the subject's speech to the start of the response. Each condition was used six times

Without Filler:
0, 500, 1000
1500, 2000, 2500, 3000 ms

With Filler:
1500, 2000, 2500, 3000 ms

By learning the natural lengths of transition intervals for all conditions of each context (e.g., the types of dialog acts of the preceding speaker), the system will successfully present a Visual Filler at an appropriate timing after the preceding speaker attempts to release the floor and sustains it until the arrival of the succeeding speaker's response. Moreover, since the Visual Filler uses the visual modality, it does not interfere with the actual utterances in the auditory modality and can be used as a supplement.

Experiment

Experiments with a natural conversation are often difficult to analyze because they are influenced by a variety of conditions such as the topics, types, and lengths of the conversations. A simulated situation of a dyadic one-turn dialog was therefore used with the constant method in order to concentrate on evaluating the effectiveness of Visual Fillers.

Procedure

A subject was seated in front of a display and was required to wear a headset. Both the display and the headset were connected to the same computer. (1) The subject was required to speak briefly about a holiday experience to the person displayed on the screen. (2) The person on the screen responded with "Oh, yeah?" after the end of the subject's speech with a certain pause that was randomly selected from 0 to 3 s. In order to regulate the situation, we used a recorded video, whose length was automatically edited online, for this response. (3) When the selected pause was 1.5 s and over, a Visual Filler was started to present 1 s after the end of the subject's speech at a chance of 50%; the filler was disappeared as the response in the screen started. The used Visual Filler is shown in figure 4. (4) After this statement-response dialog, the subject was asked, "Do you prefer a quicker reply?" and was expected to reply with either "yes" or "no." (5) The steps from (1) to (4) were repeated 66 times in total. The condition for each trial was selected randomly from the 11 conditions shown in table 1.

Finally, the ratio of the "yes" replies to the total replies was considered as the *rate of unnatural turn-taking* for each condition.

Participants

Seven subjects were considered for the experiment. All subjects were acquainted with the person on the screen because the natural transition interval had to be shared by the subjects and the person on the screen in advance and was not to be affected by unacquaintance.

Results

By comparing the rate of unnatural turn-taking for the condition "with Visual Filler" to that for "without Visual Filler," it was observed that the rates decreased or remained constant for all conditions (i.e., at the interval lengths of 1500, 2000, 2500, and 3000 ms) for six out of seven subjects. Figure 5 (left) shows the result of one subject.

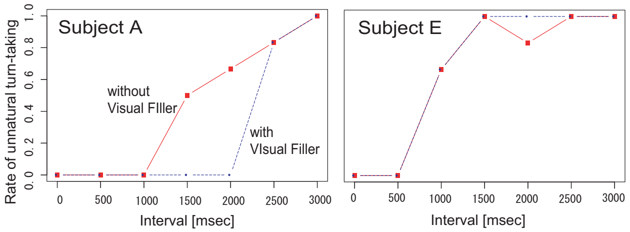

figure 5. Rate of unnatural turn-taking for each condition.

Table 2 shows the decreased rates from the condition “without Visual Filler” to “with Visual Filler.” We see that the effects of the Fillers continue for about 500-1500 ms except for one subject, and the lengths of the effects strongly depend on the person. Meanwhile, no effect of the Fillers was seen in subject E. This was because the subject felt an unnatural pause at the interval length of only 500 ms, which was earlier than the insertion of the Filler, as shown in figure 5 (right).

table 2. Decreased rates of unnatural turn-taking due to the Visual Filler (underlined values are in the improved conditions).

Subject	Interval length [ms]			
	1500	2000	2500	3000
A	0.50	0.67	0	0
B	0.33	0.50	0	0
C	0	0	0.83	0.50
D	0.33	0.33	0	0
E	0	-0.17	0	0
F	0	0.67	1.00	0.50
G	1.00	0	0	0

Discussions

Although the above result was obtained from a small preliminary experiment, we see that Visual Fillers have the ability to fill redundant gaps for most of the subjects. However, the natural lengths of the transition interval differ from person to person, and the current condition of Filler insertion (1 s after the end of the speech) seems to be too long for one subject to experience smooth turn-taking. For future work, it is therefore necessary to examine additional conditions for the Filler insertion timing. Furthermore, the following issues should be investigated to incorporate Visual Fillers into practical video-conferencing systems.

Issues for a practical system

Design of visual stimuli: We also examined other types of stimuli such as changing the blinking or rotating speed of an icon beside a participant’s images, and verified that those stimuli have similar effects on filling the temporal gaps. However, all the examined stimuli were moving icons and not natural movements; these icons may be a disturbance to the participants. We are currently trying to artificially generate the changes in a participant’s head pose on the screen by exploiting video-based texturing techniques.

Insertion timing of Visual Fillers: The appropriate timing for displaying the Visual Fillers is determined by factors such as conversation contexts and the personality of participants. Therefore, the appropriate timing should be statistically learned and tuned for each participant beforehand. As mentioned earlier, the type of dialog acts may also affect this appropriate timing; for instance, a preceding speaker can endure longer temporal gaps for question-answer pairs than for statement-reply pairs. In order to automatically recognize the types of dialog acts, we can use some studies including [14].

Automatic determination of floor release: The most significant problem in applying Visual Fillers to practical video-conferencing systems is that the method requires determining the end of the preceding speaker’s speech or turn. Although the problem is challenging, there exist some studies about automatic recognition of the end of turns and generation of the system responses by using verbal and nonverbal cues (e.g., pitch and intensity patterns) [5,15], which would serve as a clue for the practical implementation of Visual Fillers.

Acknowledgements

This work is in part supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of 18049046.

References

[1] Chen, M. Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconference. In *Proc. CHI 2002*, ACM Press (2002), pp. 49-56.

[2] Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. Meeting Recorder Project: Dialog Act Labeling Guide, *ICSI TR-04-002* (2004).

[3] Duncan, S.J., and Fiske, D.W. Face-to-Face Interaction, Lawrence Erlbaum (1977), Chapter 11.

[4] Fischer, K. and Tenbrink, T. Video conferencing in a transregional research cooperation: Turn-taking in a new medium. In *Connecting Perspectives*, Shaker (2003), pp. 89-104.

[5] Fujie, S., Fukushima, K., and Kobayashi, T. Back-channel Feedback Generation Using Linguistic and Nonlinguistic Information and Its Application to Spoken Dialogue System, In *Proc. EUROSPEECH* (2005), pp. 889-892.

[6] Grayson, D.M. and Monk, A.F. Are You Looking at Me? Eye Contact and Desktop Video Conferencing. *Trans. CHI*, Vol.10, No.3, ACM Press (2003), pp. 221-243.

[7] Jouppi, N.P. First Steps Towards Mutually-Immersive Mobile Telepresence. In *Proc. CSCW 2002*, ACM Press (2002), pp. 354-363.

[8] Kauff, P. and Schreer, O. An Immersive 3D Video-conferencing system Using Shared Virtual Team User Environments, In *Proc. CVE 2002*, ACM Press (2002), pp. 105-112.

[9] Kendon, A. Some function of gaze-direction in social interaction, *Acta Psychologica*, Vol.26 (1967), pp. 22-63.

[10] Nagaoka, C., Komori, M., Nakamura, T., and Draguna, M.R. Effects of Receptive Listening on the Congruence of Speakers' Response Latencies in Dialogues, *Psychological Reports*, Vol.97 (2005), pp. 265-274.

[11] Nguyen, D. and Canny, J. MultiView: Improving Trust in Group Video Conferencing Through Spatial Faithfulness. In *Proc. CHI 2007*, ACM Press (2007), pp. 1465-1474.

[12] Sacks, H., Schegloff, E.A., and Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language*, Vol.50 (1974), pp. 696-735.

[13] Shriberg, E.E. Spontaneous Speech: How People Really Talk, and Why Engineers Should Care. In *Proc. Eurospeech* (2005), pp. 1781-1784.

[14] Stolcke, A., et al. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech, *Computational Linguistics*, Vol. 26, No. 3 (2000), pp. 339-371.

[15] Takeuchi, M., Kitaoka, N., and Nakagawa, S. Generation of Natural Response Timing Using Decision Tree Based on Prosodic and Linguistic Information, In *Proc. EUROSPEECH* (2003), pp. 609-612.

[16] Vertegaal, R., Weevers, I., Sohn, C., and Cheung, C. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. In *Proc. CHI 200*3, ACM Press (2003), pp. 521-528.