Influence of Voice Delay on Human Perception of Group Synchronization Error for Remote Learning

One-way Communication Case

Hay Mar Mo Mo Lwin
Department of Hardware
University of Computer Studies
Yangon, Myanmar
haymarmomolwin@ucsy.edu.mm

Yutaka Ishibashi
Department of Computer Science
Nagoya Institute of Technology
Nagoya, Japan
ishibasi@nitech.ac.jp

Khin Than Mya
Department of Hardware
University of Computer Studies
Yangon, Myanmar
khinthanmya@ucsy.edu.mm

Abstract

This paper clarifies the influence of the voice delay on human perception of group synchronization error in remote learning by subjective assessment. In our assessment, we produce the difference in output timing of lecture video between a teacher terminal and a student terminal. Each subject as a student observes the output timing relation (i.e., group synchronization) between the lecture video and teacher's voice, and then the subject answers whether he/she perceives the group synchronization error or not. We also examine the influence of difference in voice contents on the error. Assessment results demonstrate that the perception rate depends on the group synchronization error plus the voice delay, and all the subjects can hardly perceive group synchronization errors (plus voice delays) less than or equal to about 250 ms. We further illustrate that the perception rate depends on the voice contents.

Keywords—*Remote learning, Voice, Video, Group synchronization error, Human perception, Subjective assessment*

I. INTRODUCTION

In multimedia applications such as video conferencing, remote learning, and networked games, the temporal relationships among multiple media streams (for example, voice and video) are important [1]-[3]. Such applications sometimes need to synchronize the output timings of several media streams at all the terminals. If the output timings of each *media unit* (*MU*) [4], which is an information unit such as a video picture and a voice packet for media synchronization, among multiple terminals are different from each other (that is, there exist group synchronization errors), the quality of experience (QoE) [5] may be damaged seriously [6]. This may affect the learning effect in remote learning, for example.

To solve the problem, we need to carry out group (or inter-destination) synchronization control [6]-[10], which adjusts the output timing of media streams among multiple terminals (or destinations). In [11] and [12], two error ranges are employed: One is the *imperceptible range* in which the synchronization error cannot be perceived by users, and the other is the *allowable range* in which the error is felt to be allowable for users. In [15], we carry out the subjective assessment to clarify the allowable and imperceptible ranges of group synchronization error. However, the voice delay is assumed to be negligible small. Also, we handle only a set of voice and video contents. We need to clarify influences of voice delay and voice and video contents on the group synchronization errors. This is because the imperceptible and allowable ranges have not been clarified so far.

Therefore, in this paper, we carry out the subjective assessment of group synchronization error by changing the voice delay to investigate the influence of voice delay in remote learning. We also examine the influence of voice contents on the error.

The rest of this paper is organized as follows. We give an outline of the group synchronization in remote learning in Section 2. We also explain the assessment method in Section 3. Then, assessment results are presented in Section 4. Finally, Section 5 concludes the paper.

II. GROUP SYNCHRONIZATION IN REMOTE LEARNING

In multicast communications, we need to perform group synchronization control, which tries to output each MU simultaneously at all the different terminals. If the control is not carried out, the MU cannot be outputted at the same time at the terminals; that is, the group synchronization error occurs.

The configuration of our remote learning system is shown in Fig. 1. The system consists of N ($\geqq 1$) terminals (one terminal is for a teacher, and the other terminals are for students) [15]. The teacher terminal uses a microphone, and each student

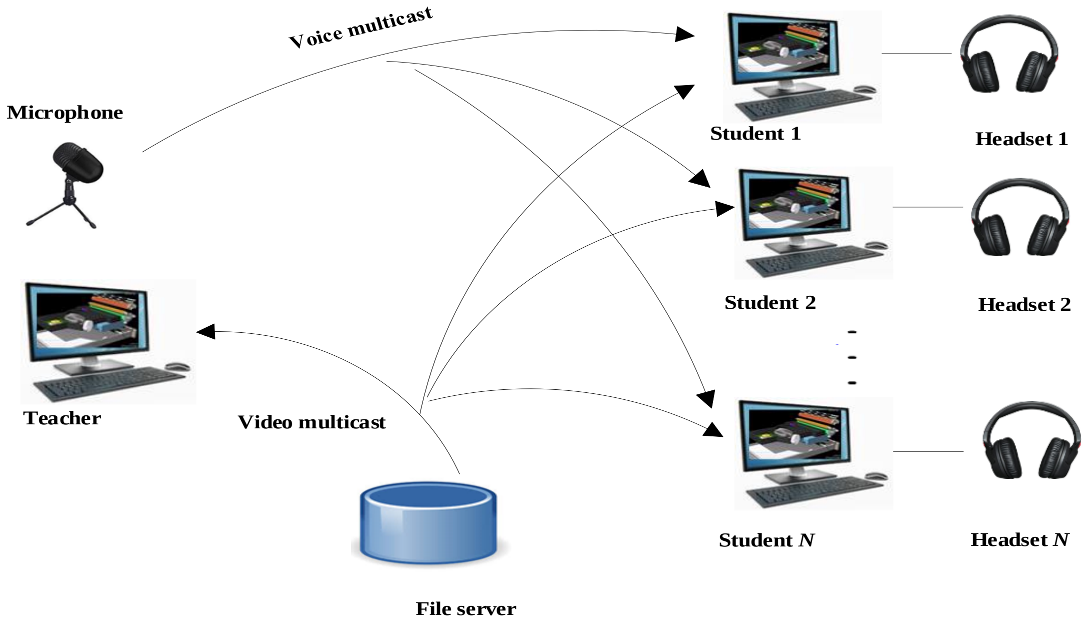

Figure 1. System configuration of remote learning.

terminals employ a headset. The file server multicasts a video stream to all the teacher and student terminals. The teacher orally explains the video contents while watching the video. A voice stream of the teacher captured via the microphone is multicast from the teacher terminal to all the student terminals. Each student listens to the teacher's voice while watching the same video [15].

III. ASSESSMENT METHOD

In the assessment, we set N=1 for simplicity as shown in Fig. 2, where two terminals were used; one was for a teacher, and the other was for a student. For simplicity, the two terminals were located at different places in a room so that the student (teacher) could not see the screen of the teacher (student) terminal. We did not use the microphone at the teacher terminal, but we used the headset for the student terminal.

Alternatively, we saved the teacher's oral explanation as a voice file at the teacher and student terminals, and the corresponding video file was also saved at the two terminals. We produced the group synchronization error of video by changing the start time of video output at the two terminals. The student terminal started to output the voice file after the voice delay when the teacher terminal started to output the voice and video files. By these setups, we constructed our assessment environment without transmitting voice and video between the two terminals. We used the video and voice contents explaining I/O devices such as a mouse and a printer in remote learning, and their output duration was 1 minute and 32 seconds [15].

At the beginning of assessment, we presented a situation in which the group synchronization error is zero and the voice delay is zero to each subject at the student terminal; that is, we started to output the voice and video files simultaneously at the student terminal. We used the single-stimulus method [13].

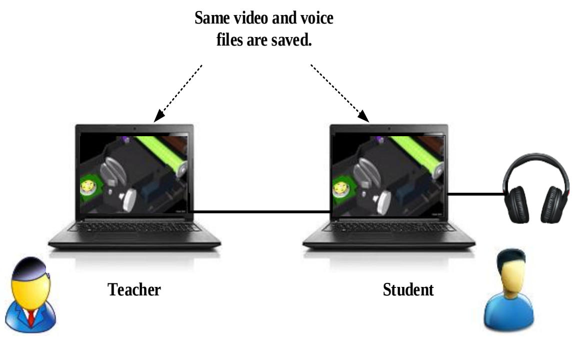

Figure 2. Assessment system

During the assessment, each subject watched the lecture video and listened to the teacher's voice. We changed the group synchronization error from –700 ms to +550 ms at intervals of 50 ms, where negative values denote the teacher ahead of the student and positive values do the teacher behind of the student, and we also changed the voice delay from 0 ms to 100 ms at intervals of 50 ms. Combinations of the error and delay were presented to each subject in random order. To examine the influence of the voice delay more clearly, we also changed the error from –100 ms to 100 ms and the delay from 0 ms to 600 ms at intervals of 50 ms. The subject did not know the value of the error and delay presented in the assessment.

After presenting each combination of the error and delay, we asked the question to each subject (student) as follows: "Did you perceive the group synchronization error?" The subject answered either "Yes" or "No." He/she judged whether the error was perceived or not by monitoring the temporal relation between the teacher's voice and displayed video images.

The total assessment time per subject was about 95 minutes including the explanation time of how to judge the error. The number of subjects (students) was 15 females, and their ages were between 28 and 37.

IV. ASSESSMENT RESULTS

We plot the *perception rate* which is here defined as the percentage of subjects who perceived group synchronization errors as a function of the group synchronization error in Fig. 3, where we set the voice delay is 0 ms, 50 ms, and 100 ms. We also show the perception rate versus the voice delay for group synchronization errors from –100 ms to 100 ms in Fig. 4.

In Fig. 3, we see that the perception rate is 0% when the group synchronization error is between around –250 ms and 250 ms and the voice delay value is 0 ms. As the absolute error exceeds about 250 ms, the perception rate starts to increase; when the absolute error is larger than or equal to about 500 ms, the perception rate is 100%. We also find in the figure that when the delay is 50 ms and 100 ms, we can obtain the perception rate by shifting the results at the delay of 0 ms leftward (i.e., the negative direction) by 50 ms and 100 ms, respectively.

In Fig. 4, we notice that the perception rate starts to increase when the voice delay exceeds about 250 ms and the group synchronization error is 0 ms; for the other errors, we can obtain the perception rates by shifting the rate at the error of 0 ms by the error. For example, the rate at the error of 100 ms is gotten by shifting that at the error of 0 ms leftward by 100 ms.

According to the above discussions, we may be able to say that the perception rate depends on the group synchronization error plus the voice delay. To confirm this, we plot the perception rate versus the error + delay in Fig. 5, which includes the perception rate versus the error (the delay: 0 ms) for different voice contents from those in this paper (called the latter *Voice 1* and the former *Voice 2* here. We will explain the results later). The figure reveals that the perception rate is almost the same as that at the voice delay of 0 ms in Fig. 3. Therefore, we can conclude that the perception rate depends on the group synchronization error + the voice delay.

From Fig. 5, we can also know the imperceptible range and the allowable range. If we assume that the imperceptible range is a range in which the perception rate is less than 20%, the range is between around –250 ms and 250 ms. If the allowable range is assumed to be a range in which the perception rate is greater than or equal to 60% (the value of 60% is just an example), the range of the absolute error is larger than about 300ms.

Furthermore, in Fig. 5, the shape of the perception rate seems to be almost line-symmetric. To confirm the line-symmetric property, we checked whether there are significant differences between the positive part and the negative part in the figure by carrying out t-test. As a result, the p-value were greater than 0.05 [14]. Therefore, we can say that the line symmetry property exists in the figure.

In addition, to examine the influence of difference in voice contents, we show the perception rate in Fig. 5 when we used the voice contents in [15] (i.e., Voice 2) and the voice delay was 0 ms. Note that Voice 2 also explains the video contents handled in this paper. The number of words in Voice 1 is about 110, and that in Voice 2 is around 170; that is, Voice 1 is simplified from Voice 2 by reducing the number of words, and Voice 1 has longer silence periods than Voice 2. From Fig. 5, we notice that the perception rate for Voice 2 is 0% when the group synchronization

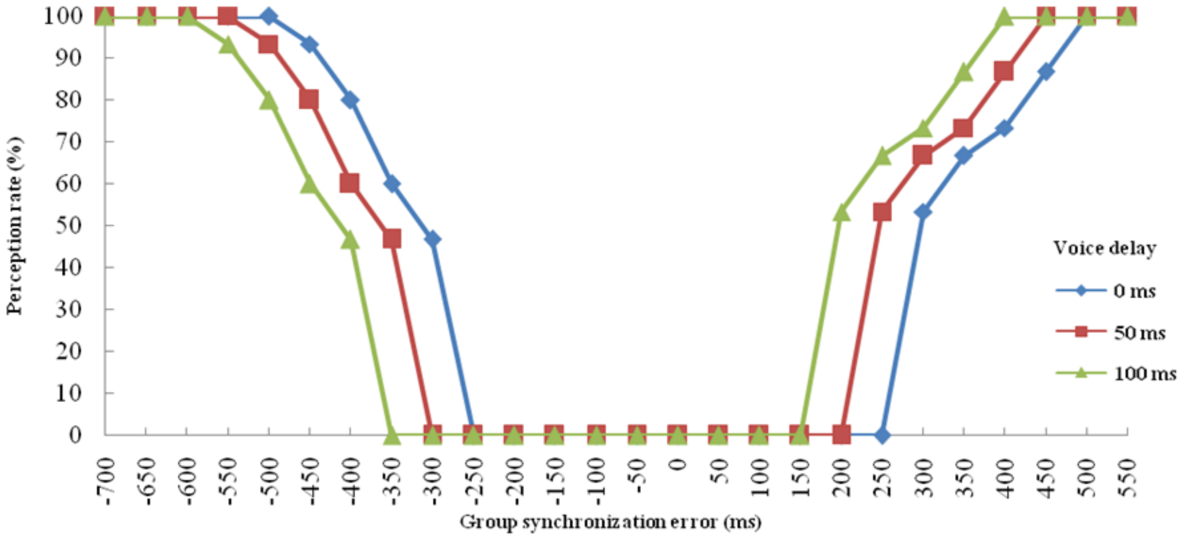

Figure 3. Perception rate versus group synchronization error.

error is between –150 ms and 150 ms (these results are almost the same as those in [15]); as we described earlier, the perception rate for Voice 1 is 0% when the error is between –250 ms and 250 ms. That is, the imperceptible range of Voice 1 is different from that of Voice 2 (on the other hand, the allowable ranges of Voices 1 and 2 are almost the same as each other). Therefore, we can conclude that the perception rate of group synchronization error depends on the voice contents. We need to clarify the influence of the voice and video contents on the perception rate; this is for further study.

V. CONCLUSION

In this paper, we investigated the influence of voice delay on human perception of group synchronization error for remote learning by carry out subjective assessment. Assessment results showed that the perception rate depends on the group synchronization error + the voice delay, and the perception rate is 0% when the error + delay is from about –250 ms to 250 ms. We also confirmed that the perception rate is dependent on the voice contents.

As our future studies, we will carry out the assessment with other video and voice contents. We will investigate the two-way communication case in which a teacher and students can interactively discuss the lecture.

REFERENCES

[1] T. S. Georgiev and E. Georgieva, “A business English multimedia distance learning application: English for tourism,” in Proc. International Scientific Conference CompSysTech, June 2002.

[2] A. Guercio, T. Arndt,andS. K. Chang, “A visual editor for multimedia application development,” in Proc. The 22nd International Conference on Distributed Computing Systems Workshops, July2002**.**

[3] A. VakiliandJ.C. Gregoire, “QoE management in a video conferencing Application,” Computer Networks, The International Journal of Computer and Telecommunications Networking, vol. 57, issue 7, pp. 1726-1738, May2013.

[4] Y. Ishibashi and S. Tasaka, “A synchronization mechanism for continuous media in multimedia communications,” in Proc. IEEE INFOCOM, pp. 1010–1019, Apr. 1995.

[5] ITU-T Recommendation G.1070, “Opinion model for video-telephony applications,” International Telecommunication Union, June2018.

[6] Y. Miyashita, Y. Ishibashi, N. Fukushima, S. Sugawara, and K. E. Psannis, “QoE assessment of group synchronization in networked chorus with voice and video,” in Proc. IEEE TENCON, pp. 393-397, Nov. 2011.

[7] Y. Ida, Y. Ishibashi, N. Fukushima, and S. Sugawara, “QoE assessment of interactivity and fairness in first person shooting with group synchronization control,” in Proc. The 9th Annual

Workshop on Network and Systems Support for Games (NetGames), Nov. 2010.

[8] K. Hosoya, Y. Ishibashi, S. Sugawara, and K. E. Psannis, “Group synchronization control considering difference of conversation roles,” in Proc. The 13th IEEE International Symposium on Consumer Electronics (ISCE), pp. 948-952, May 2009.

[9] K. Hosoya, Y. Ishibashi, S. Sugawara, and K. E. Psannis, “Effects of group synchronization control in networked virtual environments with avatars,” in Proc. The 12th IEEE/ACM International Symposium on Distributed Simulation andReal-Time Applications (DS-RT), pp. 119-127, Oct. 2008.

[10] Y. Ishibashi, M. Nagasaka, and N. Fujiyoshi, “Subjective assessment of fairness among users in multipoint communications,” in Proc. ACM SIGCHI ACE, June 2006.

[11]P. Huang, Y. Ishibashi, and M. Sithu, “Enhancement of simultaneous output-timing control with human perception of synchronization errors among multiple destinations,” in Proc. The 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 2099-2103, Oct. 2016.

[12] Y. Ishibashi, T. Kanbara, and S. Tasaka, “Inter-stream synchronization between haptic media and voice in collaborative virtual environments,” in Proc. ACM Multimedia, pp. 604-611, Oct. 2004.

[13] ITU-R Recommendation BT.500, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Sep. 2009.

[14] D. Figueiredo, “When is statistical significance not significant?”Brazilian Political Science Review, vol. 7, no. 1, Jan.2013.

[15] H. M. Mo Mo Lwin, Y. Ishibashi, and K. T. Mya, “Human perception of group synchronization error for remote learning: One-way communication case,” in Proc. IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), May2019.

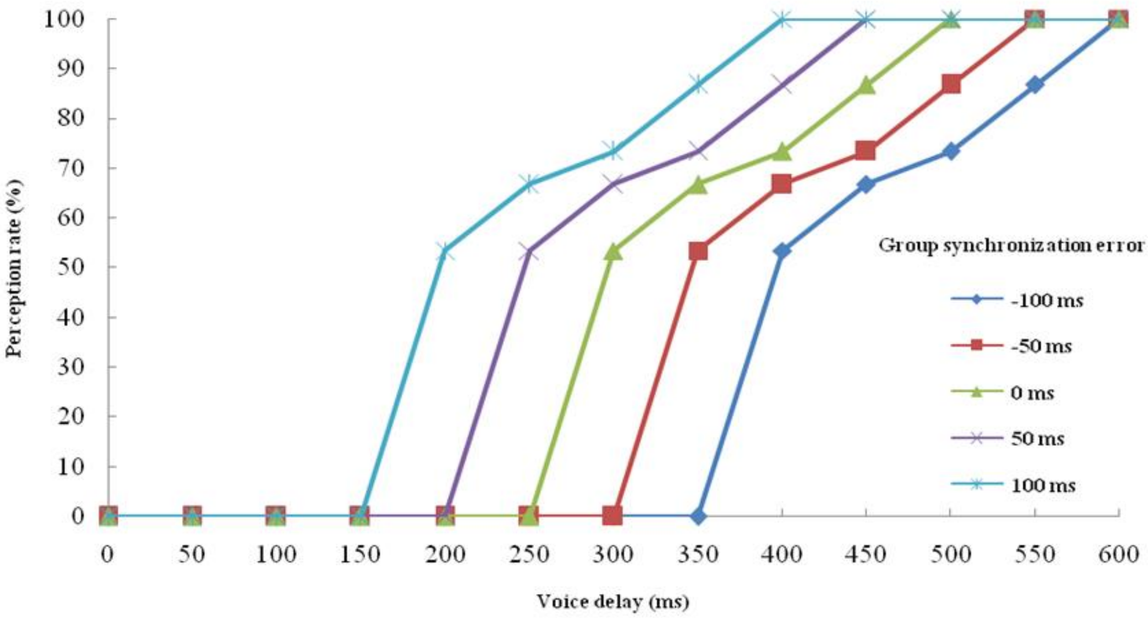

Figure 4. Perception rate versus voice delay.

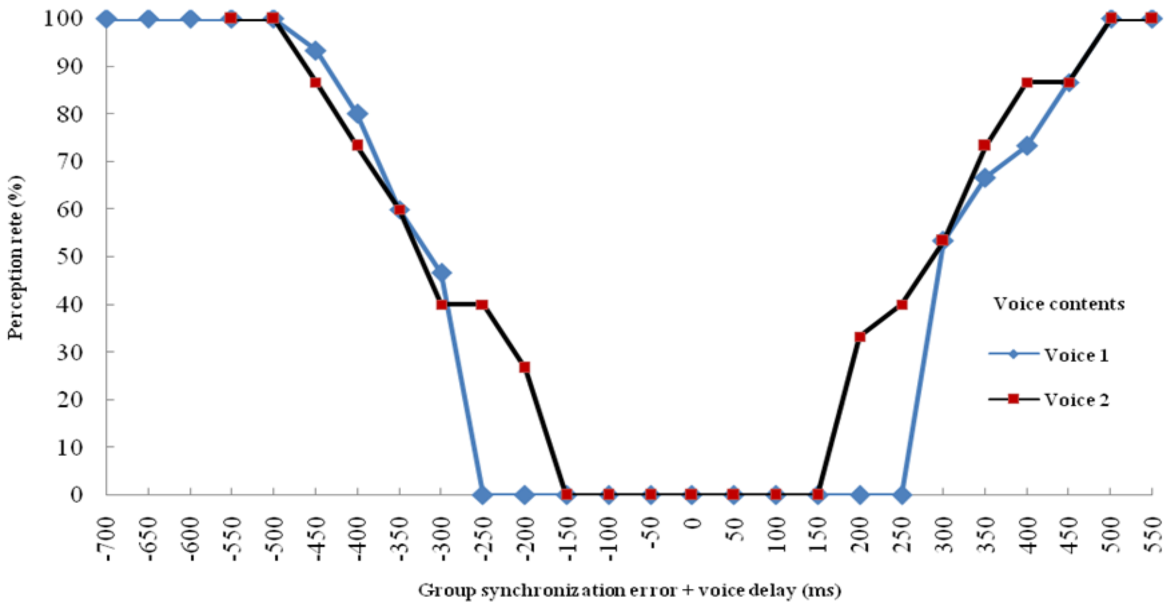

Figure 5. Perception rate versus group synchronization error + voice delay.