Effects of Automated Transcription Delay on Non-native Speakers' Comprehension in Real-time Computer-mediated Communication

Lin Yao[1], Ying-xin Pan[2], and Dan-ning Jiang[2]

[1] Institute of Psychology, Chinese Academy of Sciences, 10A Datun Road, Beijing, China
[2] IBM China Research Lab, Beijing, China
yaol@psych.ac.cn and {panyingx, jiangdn}@cn.ibm.com

Abstract. Real-time transcription generated by automated speech recognition (ASR) technologies with a reasonably high accuracy has been demonstrated to be valuable in facilitating non-native speakers' comprehension in real-time communication. Besides errors, time delay often exists due to technical problems in automated transcription as well. This study focuses on how the time delay of transcription impacts non-native speakers' comprehension performance and user experience. The experiment design simulated a one-way computer-mediated communication scenario, where comprehension performance and user experiences in 3 transcription conditions (no transcript; perfect transcripts with a 2-second delay; and transcripts with a 10% word-error-rate and a 2-second delay) were compared. The results showed that the participants can benefit from the transcription with a 2-second time delay, as their comprehension performance in this condition was improved compared with the no-transcript condition. However, the transcription presented with delay was found to have negative effects on user experience. In the final part of the paper, implications for further system development and design are discussed.

Keywords: Real-time transcription, Delay, Comprehension performance, User experience

1 Introduction

Globalization is driving more and more people to communicate using their non-native language via audio/video conferences. However, as studies have indicated, understanding the speech of a second language often poses many difficulties [7]. Thus, non-native speakers frequently find it difficult to follow the conference and the collaboration tends to be ineffective.

Pan et al. [4] proposed an approach of using real-time speech transcription to improve non-native speakers' comprehension in long-distance communications, and further explored the possibility of using automatically generated transcription [5]. Their results indicated that when synchronized with the audio stream, the automated transcripts with a word error rate (WER) of 10% could significantly improve non-native speakers' comprehension, while transcripts with a WER greater than 20% would lead to no improvement in comprehension.

Unfortunately, in addition to recognition errors, time delay often exists in automated transcripts as well. The delay primarily results from the processing time the ASR system takes. Even using the most advance speech recognition technology, the ASR processing can still be behind the audio stream for complex recognition tasks or low-quality signals. For a distributed ASR service hosting system like the one described in [2], the ASR processing will slow down when many concurrent users request the ASR service at the same time. The network delay for data transmission between the speech recognition client, server, and text showing interface also contributes to the overall delay. This study will investigate how time delay affects the usefulness of automated transcription in improving non-native speakers' comprehension in real-time communication.

Most previous studies on the effects of transcription delay have focused on helping people with hearing impairment to better understand audio or video contents [1,3,8]. Burnham et al. [1] and Maruyama et al. [3] examined hearing-impaired people's enjoyment and intelligibility of TV when captions delayed from 0 second to 4 seconds. They reported that both enjoyment and intelligibility diminished when the delay existed, and the permissible limit for delay varied from 1.63 to 4.84 seconds depending on the degree of hearing impairment and caption formats. Zekveld et al. [8] measured the benefits of transcription with speech reception threshold (SRT), and reported that delaying the transcription with 2, 4, or 6 seconds reduced the benefits of transcription by approximately 1 to 2 dB of SRT.

While these findings provided valuable insights into how time delay affects the usefulness of transcription, none of them touched upon using automated transcription to improve non-native speakers' comprehension. Furthermore, as non-native speakers need extra cognitive efforts to process the transcripts in a second language as one additional source of information, the delay could distract attention and result in little value of transcripts.

In this paper, we report an experiment that examines the effects of time delay of transcription on non-native speakers' comprehension performance and user experience, in which two research questions are addressed:

- Does automated transcription still help non-native speakers' comprehension in computer-medicated communication when a reasonable level of time delay exists in the transcripts?
- How does the time delay of transcription affect non-native sparkers' user experience?

2 Method

2.1 Preliminary Study

Before the main experiment, we did a preliminary study to find out a time delay level worth being studied more thoroughly. We started from 2 seconds and 4 seconds, which, according to previous studies [1,3,8], might be a critical level of time delay for the transcription to be usefully and acceptable. 12 Chinese participants were divided

into 2 groups and were asked to watch 6 English clips in 3 conditions: no transcript was displayed (NT), transcripts with no delay (PT) and delayed transcripts (DT). Transcription delay of the two groups was set as 2 seconds and 4 seconds respectively. After each clip was played, the participants were asked to answer 5 comprehension questions to evaluate how well they understood the materials.

The results showed that when delay was 4 seconds, the transcripts did not help the comprehension. The comprehension score of using the delayed transcripts was even worse than that when no transcript was displayed. All participants reported that they felt really frustrated by the delay and preferred to just ignoring the transcripts. In contrast, when the transcription delay was 2 seconds, the comprehension performance was improved compared to the no transcript condition, though some of the participants still reported that the delayed transcripts were somewhat distracting. Thus, in the formal experiment, we will use 2 seconds delay to confirm the usefulness of delayed transcripts.

2.2 Experiment Setup

Similar to [4,5], we designed a one-way computer-mediated communication (CMC) scenario, in which native English speakers talked in English via an audio and video channel, and native Chinese "listeners" (the participants) tried to understand what was spoken. Though communication in this study was dominated by one or a few main speakers and others just listen, the findings or conclusions were believed to serve as a useful reference for future research on more interactive scenarios.

Figure 1 showed an example of the interface developed for the experiment. Transcripts were displayed in a streaming mode, appearing letter by letter from bottom left to right. This display mode is necessary in real-time scenarios as the speakers' words cannot be foreseen before being spoken.

Fig. 1. An interface example of the experiment design.

2.3 Experiment Design

The formal experiment was designed as a within-subject study in which participants were exposed to three different Transcription Conditions:

- NT: No transcript was displayed (the baseline case).
- DT: Transcripts with 2 seconds' delay were displayed. No error was included in the transcripts.
- D-ET: Transcripts with 2 seconds' delay and 10% WER were displayed. This condition was to examine if automated transcripts with errors could help when they were not synchronized with the audio stream. The 10% WER level was selected because it was the best accuracy that could be achieved in practice (with high-quality signal, in-domain language model, and native accent) and thus might serve as the benchmark for the most tolerable level of time delay.

2.4 Participants

Thirty highly motivated university or graduate school students from various disciplines were recruited as participants. They were non-English major native Chinese speakers and had passed CET-6 (College English Test Band 6), a national English test which is mandatory for all Chinese students if they are to get a master's degree. A curious observation, however, is that though CET-6 indicates a relatively high level of English proficiency of Chinese students, there is no guarantee that those who have passed the test can understand spoken English conversations well. The participants were of a mixed gender (16 females and 14 males), and with an average age of 23.8 years (SD=4.5, range from 20-28).

2.5 Materials and Task

Six English video clips were created, 2 for each transcription condition (NT, DT and D-ET). The clips were 3.5-minute-long on average, and covered a broad range of general topics (e.g. advertising, environmental protection, obesity, etc.). 3 clips were dialogues cut from an English TV show, and the other 3 were lectures recorded with invited foreigners as speakers. 5 comprehension questions were designed for each clip, including both short-answer questions and multiple-choice questions. All the materials had been validated in our previous research and their difficulty level was appropriate for the Chinese participants [5].

The whole experiment was computer-based. A Latin square design was implemented to counterbalance order effects. Each participant was asked to watch the 6 clips. After each clip was played, the screen turned to the question-answer page immediately and no transcript could be seen any more. After finishing the comprehension test in each Transcription Condition, the participants were asked to complete a follow-up questionnaire on user satisfaction and cognitive load. The whole procedure of the experiment took about 60 minutes on average.

2.6 Measurements

Comprehension Performance.
Performance was measured by response accuracy, that is, how many comprehension questions were answered correctly. A perfect score in each condition was 10 (5 questions*2 clips).

User experience.
User experience was assessed by user satisfaction and user cognitive load.

User satisfaction. Participants were required to respond to three satisfaction evaluation questions on a 5-point Likert scale. The three questions were: (1) Usefulness: "I think transcription is helpful for my understanding." (2) Importance: "I think transcription is important for my understanding." (3) Preference: "I would love to have transcription next time."

Cognitive Load investigated how well human resources could be employed in task completion or problem solving. Three indicators were used:

- *Perception of task difficulty.* The participants assessed the difficulty of answering the questions by indicating their agreement with the following statements on a 5-point Likert scale: "It was difficult for me to correctly answer the comprehension questions" and "I fully understood what the clips talked about."
- *Perception of concentration difficulty* measured how well one can focus their cognitive resources on the task by asking the participants to respond to the following statement: "It was difficult for me to concentrate my attention simultaneously on the information from all sources (e.g., audio, video and transcription) ."
- *Perception of understanding interference.* The participants assessed how the time delay of the transcription might interfere with their understanding by indicating their agreement with the following statements on a 5-point Likert scale: "The time delay of transcripts distracted my attention" and "The time delay of transcripts hindered my understanding of video clips"

3 Results

All the data were submitted to SPSS 15.0 for analysis.

3.1 Comprehension Performance

The comprehension performance scores in different conditions were shown in Figure 2. A repeated measures ANOVA was used to analyze the data. The results showed that *Transcription Condition* had a significant main effect on performance, $F(2, 58) = 7.27$, $p < .01$, indicating that comprehension was indeed influenced by the transcription condition.

To further explore the difference between the comprehension performance in NT, DT, and D-ET, multiple comparisons were performed. The comprehension

performance in DT, D-ET was found to be significantly better than that in NT (both $p < .01$). Performance in DT was a little better than that in D-ET (5.34 vs. 5.07), but the difference did not reach a significant level ($p > .05$). These results suggested the usefulness of automated transcription when the time delay is less than 2 seconds.

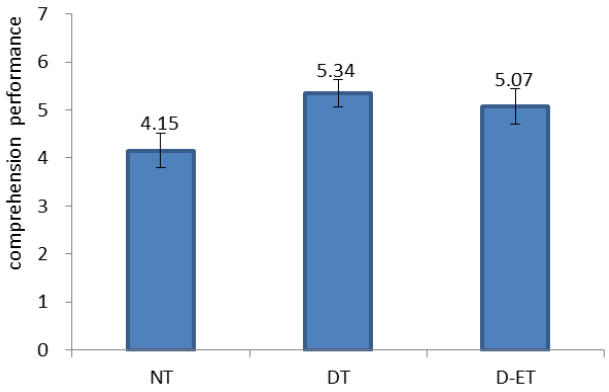

Fig. 2. Comprehension performance in NT, DT, and D-ET conditions. Error bar stands for one standard error.

3.2 User Satisfaction

The user-reported satisfaction scores for DT and D-ET condition were shown in Figure 3. The participants confirmed the usefulness of the delayed transcripts (the usefulness score in DT and D-ET condition was 4.21 and 4.03 respectively), while the importance and preference scores were nearly neutral. The results indicated that the participants did not feel very pleasant with the delayed transcripts though they were regarded as being useful

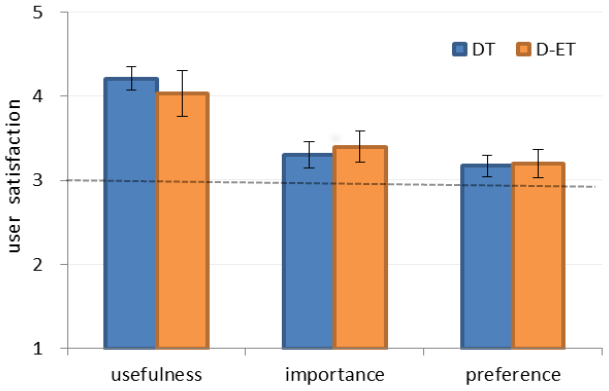

Fig. 3. User satisfaction in the DT and D-ET conditions.

3.3 Cognitive load

Cognitive load scores were presented in Table 1 in three dimensions. For task difficulty, it was found that the delayed transcripts did not increase the perceived difficulty of the comprehension test ($p >$. 05, repeated measures ANOVA). With regard to the perception of concentration difficulty and understanding interference, the results suggested a negative impact in general. The majority of the participants (66.7% in DT, 70% in D-ET,) agreed or strongly agreed that it was difficult to concentrate their attention simultaneously on the information from all sources (the means were 3.97 and 4.03 respectively). In addition, over half of the participants (56.7% in DT, 63.3% in D-ET) agreed or strongly agreed that the time delay of the transcripts would interfere with their understanding (the means were 3.80 and 3.87 respectively).

Table 1. Users' cognitive load in the NT, DT, and D-ET conditions (data were presented as means on a 5-point scale).

Cognitive load	NT	DT	D-ET
Task difficulty	2.70	2.47	2.55
Concentration difficulty	N/A	3.97	4.03
Understanding inference	N/A	3.80	3.87

4 Discussions, conclusions, and future work

In this paper, we investigated how time delay in automated transcription produced by a speech recognition system affects non-native speakers' comprehension and user experience. The results demonstrated the value of delayed transcription in improving non-native speakers' comprehension in one-way communication scenario. When the time delay was 2 seconds, the participants' comprehension performance was significantly improved with the aid of the transcripts, and the users' self-reported satisfaction also confirmed the usefulness of the transcripts. But the users' self-reported measures still showed some negative effects of time delay, e.g. the increase of concentration load, the distraction time delay has causes, and its interference with the understanding of audio information.

It seems somewhat surprising that while the non-native speakers' comprehension performance was factually improved by using the transcripts, they still reported some negative user feelings. This can be explained from several aspects. First, the task being simulating the passive one-way communication, instantaneous response on the part of the users was not required. Thus, despite the time delay, the appearance of the transcripts provided a chance for gist extraction and therefore improves the comprehension [6]. Second, the users had to pay more attention and work harder when there was time delay in the transcripts. The increased attention would result in better comprehension. But paying more attention and working harder would be more stressful and decrease the satisfaction. Third, the output delay was not only obvious enough to be perceived by users but also rendered this type of transcription very much

different from ordinary types of transcription (e.g. DVD captions) with which users are already quite familiar, hence they are inclined to make a negative assessment.

In summary, this study demonstrates that automated transcription in a good accuracy and with a reasonable level of delay (<=2 seconds) can significantly improve non-native speakers' comprehension, though user experience evaluations are not all positive. The paper implies that a certain level of transcript delay which temporally happens at system busy time (e.g. caused by large number of concurrent users) can be acceptable. But since time delay would result in negative user experience, the system should ensure that important conferences can get sufficient computation resources and high quality network connection to avoid the delay.

Future work will investigate the effects of time delay in more interactive scenarios involved in remote collaborations. In addition, finer levels of word error rate in automatically generated transcripts combined with delays should be studied, as WER in automated transcription could change within a broad range from 10% to over 30%.

Reference

1. Burnham, D., Robert-Ribes, J., and Ellison, R. Why captions have to be on time. In: Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98), pp. 153-156 (1998)
2. Jiang, D., Pan, Y., Liu, W., Qin, Y., Picheny, M., and Luther, P. Real-time Speech Transcription Service to Improve Non-native Speaker's Listening Comprehension. In: the 25th Annual International Technology & Persons with Disabilities Conference, (2009).
3. Maruyama, I., Abe Y., Sawamura E., et al. Cognitive Experiments on Timing Lag for Superimposing Closed Captions. In: Proceedings of the Sixth European Conference on Speech Communication and Technology, pp. 575-578 (1999)
4. Pan, Y., Jiang, D., Picheny, M., and Qin, Y. Effects of real-time transcription on non-native speaker's comprehension in computer-mediated communications. In: Proceedings of the 27th international conference on Human factors in computing systems, CHI 2009, pp. 2353-2356, ACM Press (2009)
5. Pan, Y., Jiang, D., Yao, L., Picheny, M., and Qin, Y. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In Proceedings of the 28th international conference on Human factors in computing systems, CHI 2010, pp. 1725-1734, ACM Press (2010)
6. Tucker, S., Kyprianou, N., and Whittaker, S. Time-Compressing Speech: ASR Transcripts Are an Effective Way to Support Gist Extraction. In: Proceedings of the 5th Joint Workshop on Machine Learning and Multimodal Interaction, pp. 226-235(2008)
7. Tyler, M.D. The Effect of Background Knowledge on First and Second Language Comprehension Difficulty. In: Proceedings of the 5th International Conference on Spoken Language Processing, (1998).
8. Zekveld, A.A., Kramer, S.E., Kessens, J.M., Vlaming, M.S., and Houtgast, T. The influence of age, hearing, and working memory on the speech comprehension benefit derived from an automatic speech recognition system. Ear and hearing, 30, 262-272 (2009).