RC24920 (C0912-012) December 11, 2009
Computer Science

IBM Research Report

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

Ying Xin Pan, Dan Ning Jiang, Yong Qin
IBM Research Division
China Research Laboratory
Building 19, Zhouguancun Software Park
8 Dongbeiwang West Road, Haidian District
Beijing, 100193
P.R.China

Michael Picheny
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

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

ABSTRACT

Real-time transcription has been shown to be valuable in facilitating non-native speakers' comprehension in real-time communication. Automated speech recognition (ASR) technology is a critical ingredient for its practical deployment. This paper presents a series of studies investigating how the quality of transcripts generated by an ASR system impacts user comprehension and subjective evaluation. Experiments are first presented comparing performance across three different transcription conditions: no transcript, a perfect transcript, and a transcript with Word Error Rate (WER) =20%. We found 20% WER was the most likely critical point for transcripts to be just acceptable and useful. Then we further examined a lower WER of 10% (a lower bound for today's state-of-the-art systems) employing the same experimental design. The results indicated that at 10% WER comprehension performance was significantly improved compared to the no-transcript condition. Finally, implications for further system development and design are discussed.

Author Keywords

Real-time Transcription, Automated Speech Recognition, Non-native Speakers, Experiment, CMC

ACM Classification Keywords

H5.1 Multimedia Information Systems, H5.2 User Interfaces

INTRODUCTION

As technology continues to facilitate collaboration across broad distances, collaborations involving people who speak different languages are becoming increasingly common. In most electronic meetings or conference calls involving a multilingual group, all members must share a common language to be able to communicate with each other. As studies have indicated, understanding speech in a second language often poses many difficulties [18]. Thus, non-native speakers frequently find it difficult to follow the meeting and the collaboration tends to be ineffective.

The most direct solution to improving non-native speakers' comprehension is speech translation. This kind of system is usually implemented as a cascade composition of speech recognition, machine translation, and text-to-speech synthesis. However, as was discussed in [10], speech translation is not a feasible solution at the moment because of the negative impact on communication caused by the combination of speech recognition errors and translation errors, as well as the high cost of developing new domains. As reported in [4], even when built with state-of-the-art components, the end-to-end performance of a simultaneous speech translation system was still not satisfactory. Only slightly over a half of the original information could be delivered to the final users.

Therefore, in [11], an alternative approach of using real-time speech transcription (captions) to help non-native speakers to achieve a better comprehension was proposed. Unlike the common display mode of instantaneous closed captions as is utilized in DVD videos, captions in real-time communications are displayed in a streaming mode. That is to say, the words appear successively as the speech stream flows forward rather than appearing as an entire line when the first word of the line is to be spoken. This display mode is necessary in real-time scenarios - the speaker's words can not be foreseen before being spoken. Pan et al. [11] demonstrated the value of real-time transcription on non-native speakers' comprehension.

In this paper, we study the possibility of employing automated speech recognition (ASR) as a cost-effective way to produce real-time captions. Though ASR performance has recently been largely improved by applying advanced acoustic training and decoding algorithms [1], it is still not perfect. For some speech recognition applications (e.g. voice interactive dialogue system, audio archive browsing, etc.), the performance is relatively insensitive to the ASR errors. But in our scenario, the quality of transcription may be critical.

In real-time communication, instantaneous and accurate understanding of the information is extremely important. When the stream of real-time transcription is shown synchronized with the audio and video, the users have to process information in smaller units and have less time to figure out the mismatch between the audio and text. Errors in the transcripts can be distracting and cause misunderstandings resulting in a degradation of comprehension performance. Also, the appearance of errors in real-time transcripts can negatively influence user satisfaction and acceptance of the use of ASR. Therefore, the impact of real-time transcription errors on non-native speakers' comprehension and user experience deserves in-depth investigation.

Recognition performance, usually measured by Word Error Rate (WER) [1], varies due to many factors, including acoustic conditions (noise level, near-field or far-field microphone, etc.), speaker characteristics (accent, speed, tone, etc.), language domain (narrow or broad), and machine processing time. Under well-controlled conditions (near-field microphone, native accent, narrow language domain), a state-of-the-art real-time speech transcription system can achieve a WER of less than 10% [2]. If the condition is less controlled (e.g. multiple speakers, more general language domain such as broadcast news or conversations, etc.), the WER can increase to 20%-25% [1]. Under even less-controlled conditions, such as meetings, lectures, or voicemail transcription tasks, where the acoustic conditions are unfavorable and the speakers and topics are diverse, the WER is often higher than 30% [16].

In this research, we introduce the use of automated speech recognition to help non-native speakers achieve better comprehension in real-time communication. We investigate the influence of the WER of the transcripts on the comprehension performance and user experience of non-native speakers. We asked the following research questions:

- How does the WER of the transcripts affect non-native speakers' comprehension in multilingual communication utilizing video conferencing system? What is the critical level of WER that can be just tolerated?
- How does the WER of the transcripts affect user satisfaction in terms of usefulness, preference and willingness to use such a feature if provided?
- How does the WER of the transcripts affect the user's cognitive load in terms of perceived comprehension difficulty and understanding interference?
- How do users perceive errors in real-time transcripts?

RELATED WORK

There are now many speech recognition applications that help people communicate and access information. It is essential to understand how speech recognition performance impacts the system's efficacy and user experience.

To the best of our knowledge, the work closest to ours has applied ASR to produce real-time transcription to help disabled students take notes in classes [6,8,20]. Leitch et al. [8] interviewed 44 students with various disabilities (e.g. medical, physical, hearing, etc.) from 8 university or college test sites to collect their feedback regarding the online transcripts produced by ASR. The interviews showed that when the text was reasonably accurate (i.e. WER<15%), most students liked using the transcripts. But when the text was not accurate enough (i.e. WER>30%), the students gave negative feedback. The major drawback of this work was the lack of quantitative analysis of the effect of ASR errors, and the absence of non-native speakers in the subject pool.

Some real-time speech recognition applications are relatively less sensitive to ASR errors. Sanders et al. [13] analyzed WER in spoken-dialogue systems. Their research on the effect of ASR accuracy revealed that on the average, task completion was possible when WER was 50% or less, and ASR accuracy appeared to have a linear correlation with successful completion of air-travel planning tasks. The high rate of task completion was partially attributable to the improved dialogue strategies for accomplishing tasks despite speech recognition errors. In contrast, in our scenario, so far there have not been any effective strategies to compensate for the negative effect brought by ASR errors.

Besides real-time applications, speech recognition has also been applied in off-line scenarios such as transcribing audio/video archives. The transcripts can be used to help users find the required information without wasting much time listening to the audio [9,15]. Startk et al. [15] studied how ASR transcript quality affected user performance of audio summarization and relevance judgment tasks with the use of a transcript-enhanced audio browser. Their data revealed that as expected, users completed tasks more rapidly and played less speech with high-quality transcripts (i.e. WER<16%). Munteanu et al. [9] investigated how the quality of transcripts affected user performance in question-answering tasks. Their major conclusions were that ASR accuracy linearly influenced both user performance and experience, and transcripts having a WER of 25% or less would be useful.

Unfortunately, while these studies provide valuable insights

[1] Word Error Rate (WER) is calculated as the percentage of incorrectly recognized words (the sum of substitutions, deletions and insertions) in the test set.

into how speech recognition error affects user performance and experience in various applications, they did not touch upon using automated transcription to improve non-native speakers' comprehension in computer-mediated communication; nor did they provide insights into what level of WER is acceptable for a transcript to be useful in a CMC interface. Thus, further research is needed to investigate the effect of real-time transcripts in various WER conditions on non-native speakers' comprehension performance.

PRELIMINARY STUDY

Before the formal experiments, we first did a preliminary study to find a WER level worth being studied more thoroughly. The materials used in the study were 6 English video clips, covering a broad range of general topics addressed by native speakers. We started from WER=20%, which was produced with the application of the IBM real-time transcription system for general purposes.

In the first pilot experiment, 6 Chinese participants were asked to watch the 6 English clips in three conditions (2 clips in each condition): no transcript was displayed, perfect transcripts were displayed, and automated transcripts (WER=20%) were displayed. After each clip was played, the participant was asked to answer 5 comprehension questions for us to evaluate how well they understood the materials. The clips and questions used here were the same with those in our formal experiments. The result suggested that when the WER was 20%, the participants' comprehension performance was better than when displaying no transcript, though worse than when displaying perfect transcripts. Nearly all participants confirmed the usefulness of automated transcription.

As the preliminary result of WER=20% looked encouraging, we continued to study an even worse WER condition. The ASR accuracy was lowered by distorting acoustic signals of the video materials (i.e. by re-recording the clips with a far-field microphone). The same transcription system was applied to transcribe the distorted signals, and the resulting WER was 35%. Another 6 Chinese participants took part in the second pilot experiment. The procedure was the same as in the first pilot experiment except that the automated transcripts had a WER of 35% instead of 20%. The result suggested that transcripts with a WER of 35% would not help with the comprehension: the performance in this condition was even worse than when displaying no transcript. Half of the participants reported that with such a high WER, the transcripts were so distracting as to impede their comprehension.

The preliminary findings gave us some insights into the effect of transcription errors. Transcription with a WER of 20% could be helpful to the non-native speakers, while a WER of 35% would impair comprehension. Thus, we will first investigate the WER=20% condition in our formal experiments to confirm its usefulness and usability.

MAIN EXPERIMENTS

Similar to [11], we designed a one-way computer-mediated communication (CMC) scenario, in which native English speakers talked in English via an audio+video channel, and native Chinese "listeners" (the participants) tried to understand what was spoken. The experiment design simulated the less interactive CMC scenario where the meeting is dominated by one or a few main speakers and the others just listen. Conclusions drawn from the one-way study can also serve as a useful reference for future research on more interactive scenarios.

EXPERIMENT 1

Experiment 1 was designed as a within-subject study in which participants were exposed to different transcription conditions in a simulated one-way communication scenario.

Independent Variables

The independent variable in this experiment was the *Transcription* condition, which had three levels:

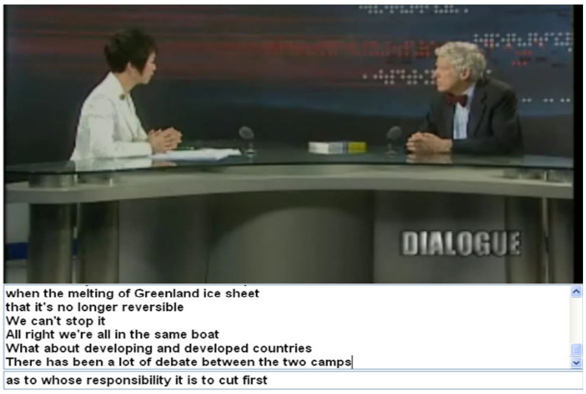
NT: no transcript was displayed (the baseline case).

PT: perfect transcripts were displayed (the ideal case).

ET-20: transcripts with errors were displayed. The transcripts had a WER of 20%, produced by a state-of-the-art general purpose speech recognition system.

Experiment Setup

The whole experiment was computer-based. Figure 1 shows an example of the interface (in this case in the PT condition) developed for the experiment. In the PT and ET-20 conditions, transcripts appeared letter by letter from bottom left to right, synchronized with the speech. All the transcripts would remain on the screen allowing participants to review if necessary. In the NT condition, the transcript display area was left blank and participants merely watched the video and listened to the audio via earphones.



when the melting of Greenland ice sheet
that it's no longer reversible
We can't stop it
All right we're all in the same boat
What about developing and developed countries
There has been a lot of debate between the two camps
as to whose responsibility it is to cut first

Figure 1. An interface example of the PT condition.

Participants

We recruited 24 university students from various disciplines as participants. They were non-English-major native Chinese speakers. All participants 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.

14 female and 10 male participated; their average age was 23.5 (SD=2.7). Before the experiment, the participants were told that their payment was dependent on how well they completed the task. By manipulating the mechanism of payment, there was a greater chance that the participants were highly motivated and would try their best to concentrate.

Materials and Task

6 English clips were created, 2 for each within-subject condition (NT, PT, and ET-20). The clips were 3.5 minutes' long on average, and covered a broad range of general topics (e.g. advertising, environmental protection, obesity, etc.) 3 clips were dialogues extracted 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 short-answer questions and multiple-choice questions.

When designing the questions, we also made sure that we balanced the number of global and local questions [17]. Global questions asked about general ideas in the text, e.g. gist, arguments (including those that can be inferred), etc., usually covering several sentences and even several paragraphs. For example, in Lecture 2, Question 5 asked about the best way control weight and the reason why it was both simple and difficult. Local questions, on the other hand, require the participants to listen for specific information, e.g. numbers, place names, etc., usually at the sentence level. For example, in Dialogue 2, Question 4 required the participants to name the three specific things that the host said were more expensive than before. All the materials had been validated in our previous research and their difficulty level was appropriate for the Chinese participants.

In formal experiments, a Latin square design was implemented to counterbalance order effects. The squares were designed such that each level of the independent variable was matched with one of the six clips appearing in any position in the sequence given to the participants.

Each participant was asked to watch 6 clips. After each clip was played, the screen shifted to the question-answer page immediately and no transcript could be seen any more. The participant was asked to answer each comprehension question within a limited time (a count-down clock was displayed on the upper left corner of the screen) and report his/her confidence level after giving each answer. After finishing the comprehension test in each *Transcription* condition, the participants were asked to complete a follow-up questionnaire on user satisfaction, cognitive load and perception of errors for the corresponding condition. The whole procedure for the experiment took about 60 minutes on average.

Measurements

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).

Confidence measured the level of confidence in the correctness of the answers. After submitting each of their answers, the participants were posed the following question: "to what extent are you confident you have given the correct answer?" (5-point likert scale)

User Satisfaction of the real-time transcription was assessed by examining how the participants responded to the following statements (5-point likert scale). (1) *Usefulness*: "I think transcription is helpful to my understanding." (2) *Preference*: "I like transcription." (3) *Willingness to use* such a feature if provided: "I would love to watch materials with transcription next time." After finishing the comprehension test in the PT or ET-20 condition, the participants were required to complete the satisfaction evaluation sheet for the corresponding condition.

Cognitive Load measured how well human cognitive resources could be employed in task completion or problem solving. The participants completed an evaluation sheet after answering all comprehension questions for each *Transcription* condition. Two indicators were used in the measurement:

- *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 understanding interference*. The participants assessed how the real-time transcription (PT or ET-20) might interfere with their understanding by indicating their agreement with the following statements on a 5-point likert scale: "The transcription distracted me" and "It was difficult for me to concentrate my attention simultaneously on the information from all sources" These statements were included only in the conditions in which transcripts were presented.

Perception of speech recognition errors. The participants were asked about their perception of speech recognition errors by indicating agreement with three statements: "I have noticed there were errors in transcripts", "The errors hindered my understanding of the clips" and "The errors devalued the usefulness of transcripts." These statements were included only for the condition in which the automated transcripts were present.

Results

All the data were submitted to SPSS14.0 for analysis.

Comprehension Performance

The comprehension performance scores with standard error bars in different conditions are shown in figure 2. A repeated measures ANOVA was used to analyze the data. The result showed that *Transcription* had a significant main effect on performance (*F* (2, 46) =10.056, *p*<.001). It indicated that user performance is indeed influenced by the transcription condition.

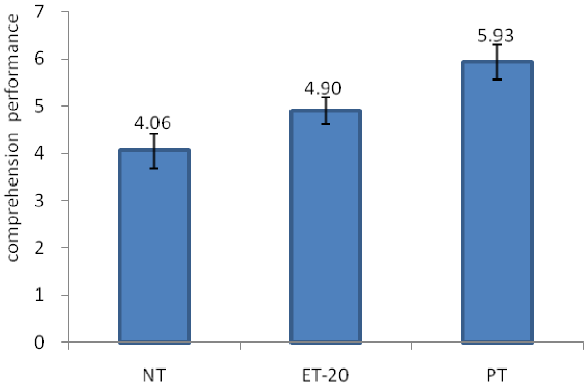


Figure 2. Comprehension performance in the NT, ET-20, and PT conditions.

To further explore the difference between the comprehension performance in NT, PT, and ET-20, we performed a set of multiple comparisons. The comprehension performance in PT was found to be significantly better than that in both NT (*t* (23) =4.465, *p* <.001) and ET-20 (*t* (23) =2.742, *p* <.01), while the performance in ET-20 was marginally better than that in NT (*t* (23) =1.847, *p* =.078). The result suggested a possibility of the usefulness of transcripts with WER=20%, though the comprehension performance in the perfect condition was still significantly better than that in the ET-20 condition.

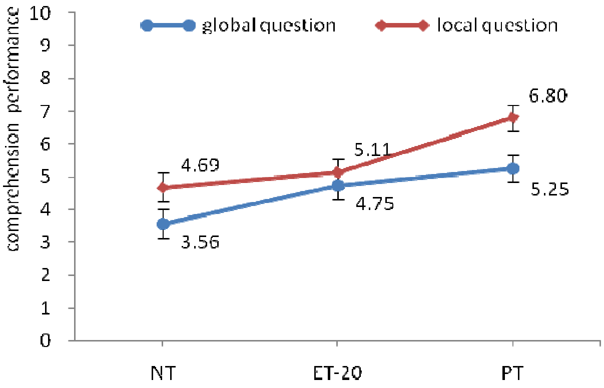


Figure 3. Comprehension performance of global and local questions in the NT, ET-20, and PT conditions.

To get more insights into the usefulness of transcripts with WER=20%, we further analyzed the performance scores relative to global and local questions respectively. As shown in figure 3, the effect of transcription performance seems different for global and local questions. Compared to the NT condition, the improvement of comprehension performance in the ET-20 condition was larger for global questions than local questions, which suggests that the comprehension of local information may be more sensitive to recognition errors. But the analysis of interaction showed that this difference did not reach a significant level (*F* (2, 46) =1.392, *p* = .259).

Comprehension Confidence

The comprehension confidence scores in different conditions are shown in figure 4. A repeated measures ANOVA showed that *Transcription* had a significant main effect on comprehension confidence (*F* (2, 46) =5.825, *p* <.01). Multiple comparisons indicated that there was a significant difference in user confidence between PT and NT (*t* (23) =2.936, *p* <.01), and between PT and ET-20 (*t* (23) =2.334, *p* < .05), while no significant difference was found between ET-20 and NT (*t* (23) =0.382, *p* =.706). The results did not suggest that transcripts with WER=20% improved the participants' confidence compared with the baseline condition NT.

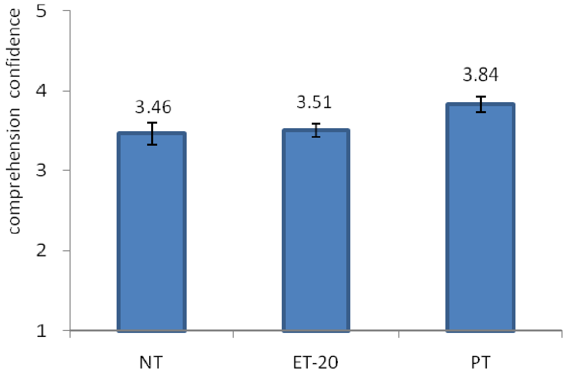


Figure 4. Comprehension confidence in the NT, ET-20, and PT conditions.

User Satisfaction

We asked the participants to report their degree of satisfaction with the use of transcription in both the PT and ET-20 conditions. The scores are shown in figure 5. The participants reported positive user satisfaction scores with the use of perfect transcripts. When the WER of the transcripts was 20%, the user satisfaction scores were still positive in all three dimensions, though lower than in the perfect transcription condition.

A paired t-test was further performed to see if the transcription errors led to a significantly different satisfaction level. PT was found to result in a significantly higher level of satisfaction score than ET-20 (*t* (23) =2.393, *p* <.05). The data suggested that though the participants reported positive satisfaction with the use of WER=20% transcripts, there was still a gap between the automated transcription and perfect transcription.

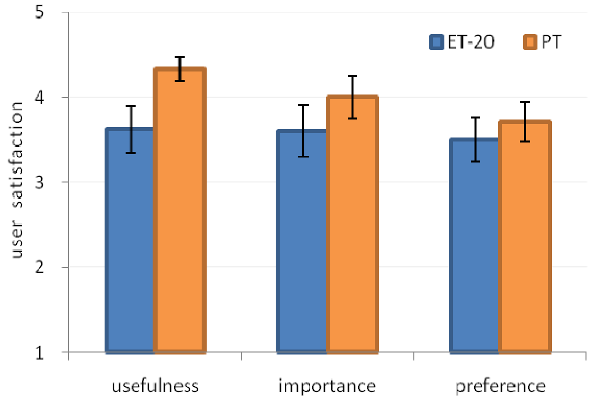


Figure 5. User satisfaction in the ET-20 and PT conditions.

Cognitive Load

Perception of task difficulty. The participants were asked to evaluate the difficulty of the task in each condition. The data are listed in table 1. A repeated measures ANOVA showed that the main effect of transcription conditions on user perceived task difficult did not reach a significant level (*F* (2, 46) =1.507, *p* =.232). The data did not suggest an increased perceived difficulty caused by errors in the transcripts.

Perception of understanding interference. The participants were asked to report their perception of the effect of speech recognition errors on understanding interference. We carried out a paired t-test to see if participants perceived a difference using the perfect and automated transcripts. The data showed that the participants did feel transcripts with WER=20% harder to understand than the perfect transcripts (*t* (23) =2.349, *p*<.05). Transcription with WER=20% were evaluated more negatively compared with the perfect transcription.

Transcription / Cognitive Load	NT	ET-20	PT
Task Difficulty	2.92	2.85	2.54
Understanding Interference	N/A	3.75	3.25

Table 1. User-perceived cognitive load in the NT, ET-20, and RT conditions.

Perception of Errors

When WER of the transcripts was 20%, 66.7% of the participants agreed with the statement "I have noticed there were errors in transcripts" and only 12.5% disagreed. It showed that the majority of participants were aware of the existence of errors in transcripts.

About the effects of errors on comprehension, the user perceptions were not positive. For the statements "The errors hindered my understanding of the clips" and "The errors devalued the usefulness of transcripts", only a small part of the participants (33.5% and 34.7% respectively) disagreed. This indicated that the errors had a negative impact on how users perceived the value of real-time transcripts.

Conclusion and Discussion

In this experiment, we investigated the effects of transcription produced by an ASR system with WER=20% on comprehension performance and user experience of non-native speakers. The result revealed that the comprehension performance in the ET-20 condition was marginally better than that in NT, but still obviously worse than that in the ideal condition of PT. To obtain more insights, we further calculated the total sum of performance scores respectively related to the global questions and related to the local questions, and analyzed the two groups of data. The data showed that the comprehension of local information might be more sensitive to recognition errors. As for comprehension confidence, we did not find any improvement using the automated transcripts.

As regards user experience, the results suggested "mixed" user feelings. The participants reported positive user satisfaction with the aid of automated transcripts, but the satisfaction level was significantly lower than that in the PT condition. Besides, though we did not find significant increase of user-perceived task difficulty caused by the transcription errors, the participants did feel transcripts with errors were harder to understand. As for user perception of errors, the majority of the participants were aware of the existence of transcription errors, and reported negative effects of errors on comprehension.

Therefore, both the comprehension data and user experience data suggested that WER=20% would be a critical level of error rate for the real-time transcription produced by an ASR system to be useful and acceptable. For any WER higher than 20%, real-time transcription would be of little use in improving non-native speakers' comprehension or in achieving a satisfactory user experience.

As the gap between the comprehension performance with WER=20% and perfect transcription was still evident, we performed another experiment (experiment 2) to investigate the condition where ASR performed the best. For experiment 2, the ASR performance was improved to WER=10% by adapting the language model with in-domain data of the materials to be transcribed, which was nearly the best accuracy that the ASR system could achieve. Thus, experiment 2 would give us an idea about to what extent the real-time transcription produced by an ASR system could help in the best-controlled circumstances.

EXPERIMENT 2

Experiment 2 was identical to experiment 1, except that WER in the automated transcripts was 10% instead of 20%. Thus, the independent variable was the *Transcription* condition with three levels:

NT: no transcript was displayed (the baseline case).

PT: perfect transcripts were displayed (the ideal case).

ET-10: transcripts with errors were displayed. The transcripts had a WER of 10%, produced by a state-of-the-art speech recognition system for a specific domain.

In both of our experiments, only one WER level of the automated transcription was studied at a time. Though it might sound more efficient to include multiple WER levels in one experiment, it nevertheless had the potential risk of overloading the non-native speakers (participants) by even more comprehension tasks and therefore failing to ensure reliable performance.

In experiment 2, we recruited another 24 Chinese university students as participants, 11 female and 13 male. Their average age was 23.7 (SD=3.3).

Results: Comprehension Performance

The comprehension performance in different conditions is shown in figure 6. Similar to the result of experiment 1, a repeated measures ANOVA indicated a significant main effect of *Transcription* (*F* (2, 46) =19.117, *p* <.001) on comprehension performance.

Multiple comparisons showed that the comprehension performance in the ET-10 condition was significantly improved compared with that in the NT condition (*t* (23) =3.448, *p* <.01). It demonstrated the value of transcripts with WER=10% in improving non-native speakers' comprehension. However, the performance in the PT condition was still significantly better than that in ET-10 (*t* (23) =2.742, *p* <.05), which showed that the improvement of comprehension with the aid of the automated transcripts (WER=10%) was still not as large as that with perfect transcription.

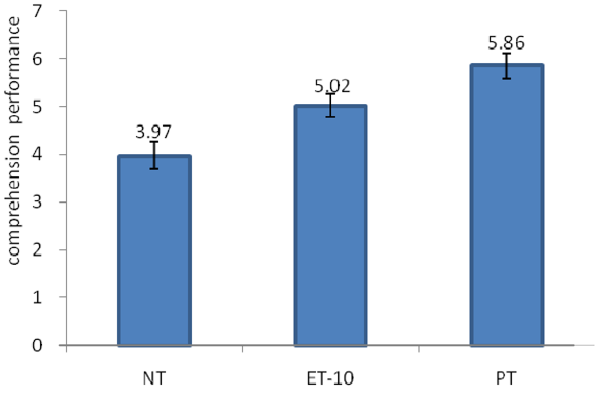


Figure 6. Comprehension performance in the NT, ET-10, and PT conditions

We further divided the comprehension performance scores into two groups, those related to global questions and those related to local questions, and analyzed the associated performance in different conditions, shown in figure 7. This time, the effect of transcription on the comprehension of global information was similar to that of local information. The comprehension of both global and local information was improved to a similar extent with the aid of the automated transcripts.

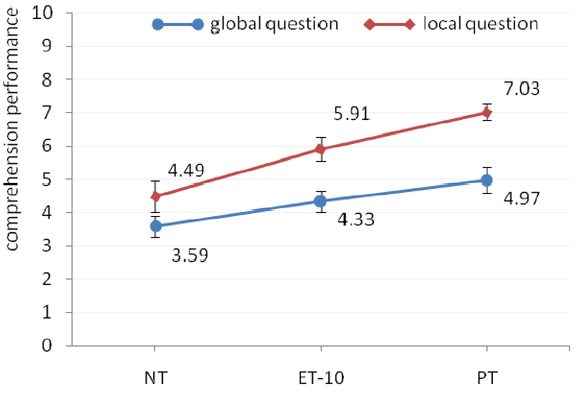


Figure 7. Comprehension performance of global and local questions in the NT, ET-10, and PT conditions.

Comprehension Confidence

The comprehension confidence scores in the NT, ET-10, and PT conditions are shown in figure 8. A significant main effect was found on the comprehension confidence in different transcription conditions (*F* (2, 46) =5.525, *p* <.01). When WER=10%, the automated transcripts (ET-10) significantly improved the confidence score compared with the NT condition (*t* (23) =5.229, *p* <.001), while no significant difference was found between the confidence scores in the ET-10 and PT conditions (*t* (23) =1.528, *p* =.137).

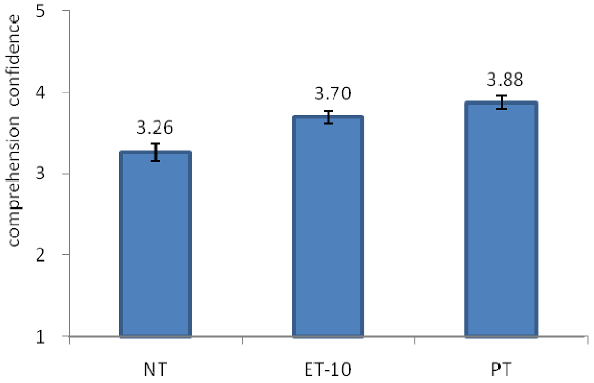


Figure 8. Comprehension confidence in the NT, ET-10, and PT conditions.

User Satisfaction

The user-reported satisfaction scores are shown in figure 9. Similar to the result in experiment 1, the participants reported positive user satisfaction in both the PT and ET-10 conditions. But when WER=10%, the paired t-test found no significant difference between the user satisfaction in the PT and ET-10 condition (*t* (23) =1.175, *p* =.252). The result suggested that the participants had a desirably positive attitude toward the transcripts when WER was reduced to 10%.

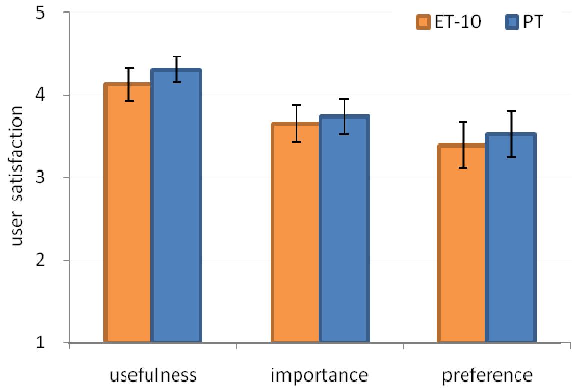


Figure 9. User satisfaction in the ET-10 and PT conditions.

Cognitive Load

Perception of task difficulty. The user-perceived cognitive load data are listed in table 2. A repeated measures ANOVA showed a significant main effect of transcription conditions on the user-perceived task difficulty (*F* (2, 46) =4.376, *p* <.05). Multiple comparisons further showed that when WER=10%, the perceived task difficulty was significantly reduced with the use of the automated transcripts (ET-10 vs. NT, *t* (23) =2.893, *p* <.05), and no significant difference was found between user-perceived task difficulty with automated transcripts and that with perfect transcripts (ET-10 vs. PT, *t* (23) =0.316, *p* =.801).

Perception of understanding interference. A paired t-test was performed to examine the level of difference between perceived understanding interference across the ET-10 and PT conditions. The results did not suggest that the perception of understanding interference was significantly enhanced when errors (WER=10%) were present in the transcripts (*t* (23) =.811, *p* =.426).

Transcription Cognitive Load	NT	ET-10	PT
Task Difficulty	3.52	2.83	2.77
Understanding Interference	N/A	3.50	3.52

Table 2. User-perceived cognitive load in the NT, ET-10, and PT conditions.

Perception of Errors

The perception of errors in the transcripts when WER=10% was very similar to the result in experiment 1. The majority of the participants (66.7%) agreed with the statement "I have noticed there were errors in the transcripts", and only a small part of them disagreed with the statements "The errors hindered my understanding" and "The errors devalued the usefulness of transcripts" (33.4% and 20.8% respectively). The data suggested that even when the WER was reduced to 10%, the participants still reported negative perception toward the appearance of errors.

Conclusion and Discussion

In experiment 2, we investigated the effects of transcription with WER=10% on non-native speakers' comprehension and user experience. The automated transcripts were produced by an ASR system in a best-controlled condition. The result showed that the participants' comprehension performance was significantly improved in the ET-10 condition compared with the NT condition, but the automated transcripts still did not help with non-native speakers' comprehension as much as the perfect transcripts. The participants' comprehension confidence was also significantly improved with the use of automated transcripts, and no significant difference was found in the ET-10 and PT conditions.

The user experience data suggested mainly positive user feelings. The participants reported positive user satisfaction levels when using the automated transcripts, and we did not find a significant difference of the satisfaction level between the ET-10 condition and the PT condition. The cognitive load data showed that the use of automated transcripts reduced the perceived task difficulty, and that there was no significant enhancement of the perception of transcription interruption caused by the transcription errors. But the user perception of errors in the transcripts was still negative.

Comparing the comprehension performance data and user experience data with those obtained in experiment 1, we found that transcripts with WER=10% worked much better than transcripts with WER=20% upon comprehension performance and user experience. However, even with such a low error rate, comprehension performance with the aid of automated transcripts was still not as good as that with perfect transcripts. The participants also reported negative feelings about errors in the transcripts. These facts suggested that the speech recognition technology still needs to be improved to better facilitate non-native speaker's comprehension.

CONCLUSIONS, DISCUSSIONS AND FUTURE WORK

In this paper, we investigated the effects of real-time transcription produced by an ASR system on non-native speakers' comprehension. We performed two formal experiments to study the relative contribution of automated transcription with a WER of 20% and 10% respectively. The transcripts with a WER of 20% were produced by a state-of-the-art ASR system for general purposes, and those with a WER of 10% were produced by the same ASR engine using the domain-specific language model, which was nearly the best accuracy level that the current speech recognition technology could achieve.

Our analysis of the comprehension performance data revealed that when WER=20%, the automated transcripts marginally improved the participants' comprehension performance, and the comprehension of local information (e.g. key words like numbers, proper nouns, etc.) might be more sensitive to recognition errors. In contrast, when the

WER was reduced to 10%, both the comprehension performance and confidence were significantly improved. The results demonstrated the value of automated transcription in our scenario, but it was also found that the performance gap between the automated transcription and perfect transcription was still evident even when WER=10%.

Regarding user experience, for both WER levels, the participants reported positive user satisfaction with the aid of automated transcription. However, the satisfaction level was found to be significantly different between the automated transcription condition (WER=20%) and perfect transcription condition, while this difference was not found to be evident when the WER was reduced to 10%. The reduction of the WER also helped to alleviate the participants' cognitive load. When the WER was 10%, no cognitive issue was reported by the participants as to the ability to synthesize different sources of information.

Though the participants gave positive feedback on the use of automated transcription, their perception of errors in the transcripts was still negative. For both WER levels, the majority of the participants were aware of the errors in the transcripts, and only a small number of them thought the errors did not hinder their comprehension or disvalue the usefulness of the transcripts. The data showed a clear user preference for perfect transcription over automated transcription.

In summary, our study demonstrated that real-time transcription produced by an ASR system can improve non-native speakers' comprehension in multilingual communication utilizing video conferencing systems. The WER of 20% tended to be a critical level at which the comprehension can be improved, and when the WER was reduced to 10%, significantly better comprehension performance and user experience can be obtained. However, we did not actually find a so-called "good-enough" level of accuracy: even with the best accuracy that could be technically achieved, the comprehension performance in the PT condition was still evidently better than when using the automated transcription, and the user perception of errors in the transcripts was negative.

The first implication of our results is that if real-time automated speech recognition is utilized to help with non-native speakers' comprehension, the WER of the transcripts *certainly* must be less than 20%, and preferably be reduced to 10% or even less for better user performance. Thus, the system needs to be carefully designed and implemented to achieve the necessary recognition accuracy. To guarantee speech signal quality, near-field microphones should be used as much as possible, and the speech recognition system should be placed at the speaker's side rather than at the listeners' side to avoid signal distortions caused by the communication channel. Given high-quality signals, a state-of-the-art speech recognition system can achieve a 20% WER when dealing with native accent and general topics. To further improve the accuracy, two types of adaptation techniques need to be applied.

Speaker adaptation techniques can be applied to better handle a certain speaker's accent and his/her personal characteristics. It can be supervised [3] or unsupervised [19, 21]. For supervised adaptation, the adaptation data are collected through a pre-training procedure in which the speaker is asked to read the prepared scripts for a few minutes. For unsupervised adaptation, the adaptation data can be collected from the speaker's previous voice data logged in the system. Besides offline adaptation, online adaptation methods can also help [2]. In some circumstances, WER can be reduced by 10%-20% relative by using speaker adaptation techniques.

Language model adaptation techniques [7] can also be applied to improve the recognition accuracy for speech of specific domains. A long-distance meeting usually has a pre-determined topic and agenda, and often has relevant presentations and documents. Such textual materials can be collected to adapt the vocabulary and language model for speech recognition. We can even exploit the online resources (the internet or the enterprise intranet) to obtain more data relevant to the meeting topic [12,14]. With an adequate adaptation of the language model, the WER could be further reduced by 15%-30% relative.

Our results also suggested that the comprehension of local information like numbers, proper nouns, etc. tended to be more sensitive to recognition errors. Fortunately, this could be compensated by the use of other presentation modalities such as PowerPoint presentations, or whiteboard handwriting [5], where the key local information is often highlighted. A future design of a conferencing system can combine real-time speech transcription with multiple presentation modalities to achieve better comprehension on the part of the non-native speakers.

Future work will study more WER levels to investigate the correlation between WER and comprehension, as well as the effects of automated transcription in two-way communication scenarios.

REFERENCES

1. Chen, S., Kingsbury, B., Mangu, L., et al. Advances in Speech Transcription at IBM under the DARPA EARS Program. *IEEE Transactions on Audio, Speech, and Language Processing 14, 5 (2006), 1596-1608.*
2. Cui, X., Gu, L., Xiang, B., et al. Developing High Performance ASR in the IBM Multilingual Speech-to-Speech Translation System. In *Proc. ICASSP 2008* (International Conference on Acoustics, Speech, and Signal Processing), IEEE Press (2008), 5121-5124.
3. Gales, M. J.F. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. *Computer Speech and Language*, 12 (1998), 75-98.

4. Hamon O, Fugen C., Mostefa D., et al. End-to-End Evaluation in Simultaneous Translation. In *Proc. 12th Conference of the European Chapter of the ACL*, 345-353.

5. Kaiser, E.C., Barthelmess, P., Erdmann, C., et al. Multimodal Redundancy across Handwriting and Speech During Computer Mediated Human-Human Interactions. In *Proc. ACM SIG'CHI 2007*, ACM Press (2007), 1009-1018.

6. Kheir, R and Way, T. Inclusion of Deaf Students in Computer Science Classes Using Real-time Speech Transcription. In *Proc. ITiCSE 2007* (Annual Conference on Innovation and Technology in Computer Science Education), ACM Press (2007), 261-265.

7. Lau, R., Rosenfeld, R., and Roukos, S. Adpative Language Modeling Using the Maximum Entropy Principle. In *Proc. the ARPA Workshop on Human Language Technology* 1993, 108-113.

8. Leith, D. and MacMilan, T. Liberated Learning Initiative Innovation Technology and Inclusion: Current Issues and Future Directions for Liberated Learning Research. *Year III Report*, 2003 Saint Mary's University, Nova Scotia.

9. Munteanu, C., Baecker, R., Penn, G., et al. The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives. In *Proc. ACM SIG'CHI 2006*, ACM Press (2006), 493-502.

10. Nakamura, S., Markov, K., Nakaiwa, H., et al. The ATR Multilingual Speech-to-Speech Translation System. *IEEE Transactions on Audio, Speech, and Language Processing 10*, 2 (2006), 365-376.

11. Pan, Y., Jiang, D., Picheny, M., et al. Effects of Real-time Transcription on Non-native Speaker's Comprehension in Computer-mediated Communications. In *Proc. ACM SIG'CHI 2009*, ACM Press (2009), 2353-2356.

12. Ramabhadran, B., Siohan, O., and Sethy, A. The IBM 2007 Speech Transcription System for European Parliamentary Speeches. In *Proc. ASRU 2007* (the Automatic Speech Recognition and Understanding Workshop), IEEE Press (2007), 472-477.

13. Sanders, G.A. and LE, A.N. Effects of Speech Recognition Accuracy on the Performance of DARPA Communicator Spoken Dialogue Systems. *International Journal of Speech Technology* 7 (2004), 293-309.

14. Shi, Q., Chu, S.M., Liu, W. et al. Search and Classification Based Language Model Adaptation. In *Proc. Interspeech 2008* (Annual Conference of the International Speech Communication Association).

15. Stark, L., Whittaker, S., and Hirschberg, J. ASR Satisficing: The Effects of ASR Accuracy on Speech Retrieval. In *Proc. of ICSLP 2000* (International Conference on Spoken Language Processing).

16. Stolcke, A., Anguera, X., Boakye, K., et al. The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System. *Lecture Notes in Computer Science*, Springer (2008).

17. Tong, K. A. The effects of question type, question position, text length and answer location on ESL listening tests. Exploring Language. Hong Kong: Language Center, Hong Kong University of Science and Technology.

18. Tyler, M.D. The Effect of Background Knowledge on First and Second Language Comprehension Difficulty. In *Proc. ICSLP 1998* (International Conference on Spoken Language Processing.)

19. Uebel, L.F. and Woodland, P.C. Speaker Adaptation Using Lattice-based MLLR. In *Proc. ITRW on Adaptation Methods for Speech Recognition*, 2001.

20. Wald, M. Using Automatic Speech Recognition to Enhance Education for All Students: Turning a Vision into Reality. In *Proc. ASEE/IEEE Frontiers in Education Conference*, S3G-22-25.

21. Woodland, P.C., Pye, D., and Gales, M.J.F. Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression. In *Proc. ICSLP 1996* (International Conference on Spoken Language Processing), 1133-1136.