REAL-TIME SPEECH TRANSCRIPTION SERVICE TO IMPROVE NON-NATIVE SPEAKERS' LISTENING COMPREHENSION

Dan-ning Jiang, Ying-xin Pan, Wen Liu, Yong Qin
IBM China Research Lab, Beijing, China
Email: {jiangdn, pangyingx, liuwen, qinyong}@cn.ibm.com

Michael Picheny
IBM Watson Research Center, Yorktown Heights, NY
Email: picheny@us.ibm.com

Paul Luther
IBM Human Ability and Accessibility Center, Austin, TX
Email: pluther@us.ibm.com

1. Introduction

Globalization is driving many people to communicate in non-native languages. As studies have indicated, listening comprehension of a non-native language poses many difficulties [1]. Thus, non-native speakers in the multilingual group frequently find the communication difficult and the collaboration tends to be ineffective. The most direct solution to improving non-native speakers' listening comprehension is speech translation [2]. The system transcribes the speech to text, translates the text into the non-native speaker's first language, and outputs speech synthesized from the translated text. However, due to technological limitations, the communication is often disrupted by the combination of recognition errors and translation errors [3]. In addition, the development of new language pairs or domains for translation is difficult and costly [2]. For both of these reasons, speech translation is not a feasible solution at the present moment.

In this presentation, we propose another approach to improving non-native speakers' listening comprehension – providing real-time transcription synchronized with the audio stream. Compared with machine translation, speech transcription technology is relatively mature [4]. Under controlled conditions (native accent, close-talking microphone, matched domain of language), a state-of-the-art English system can achieve a 5% ~ 7% Word Error Rate (WER) when transcribing conversational speech in real-time [5]. Good performance can also be obtained across various conditions (speaker, environment, domain, etc.) without much difficulty via acoustic and domain adaptation techniques.

We will first show the value of real-time transcription in improving non-native speakers' listening comprehension via a user experience study, and then present our prototype system which provides real-time speech transcription service utilizing speech recognition technology.

2. The value of real-time transcription for non-native speakers

We performed a user experience study to demonstrate the value of real-time transcription in improving non-native speakers' listening comprehension. The experiment was designed in a one-way communication scenario, in which native English speakers talked in English via audio or video channel, and native Chinese listeners (the participants) tried to understand the talks. Three

research questions were addressed:

- Does real-time transcription help non-native speakers improve listening comprehension in multilingual communication?
- How do users perceive real-time transcription in terms of usefulness, preferences, and willingness to use such a feature if provided?
- How do users allocate their cognitive resources when presented with multiple information sources?

The experiment was a 2x3 mixed design. The between-subject variable was communication *Modality*, which had two levels:

A – the communication channel was audio only
AV – the communication channel was audio+video

The within-subject variable was *Transcription*, which had three levels:

NT – no transcript was displayed
T – real-time transcript was displayed in such a way that the current line of transcript displaced the previous line of transcript
RT – real-time transcript was displayed in such a way that all records of transcription were kept and referable at any time

We recruited 48 students (both from universities and graduate schools) 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 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 English talks well.

6 English clips were made as the listening materials, 2 for each within-subject condition (NT, T and RT). 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 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.

In the experiment, Latin square design was implemented to counterbalance the order effect. Each participant was asked to listen to (*Modality* of A) or watch (*Modality* of AV) 6 clips. After each clip was played, the screen turned to the question answer page immediately and no transcript was shown any more. The participant was asked to answer each comprehension question within a limited time and report his/her confidence level after giving each answer. Finally, after all clips were finished, the participant was asked to complete a follow-up questionnaire on user experience and attention allocation.

The results of the experiment strongly demonstrate the value of transcription:

- Real-time transcription (both R and RT) significantly improved comprehension performance and confidence of non-native speakers, both in A and AV conditions.
- In addition to the improved performance and confidence, the participants also reported a

more satisfying user experience with the aid of transcription. This should not be taken for granted – there is the possibility that despite the improved performance, adding one more information source could exert so heavy a cognitive load as to render the experience unpleasant.

- The attention allocation data showed that the participants allocated more attention (about 50%) to reading the text when transcription (T or RT) was presented. This result is consistent with the comprehension result – users tend to allocate more attention to the information source that has a more direct impact on performance.

3. Prototype system

We implemented a prototype system which provides real-time transcription service to improve non-native speakers' listening comprehension. This system is in the client/server structure, as shown in figure 1. The client records the speaker's voice data, sends the voice stream to the server, and receives the transcription generated by the server. The server is composed of a cluster of distributed speech recognition engines. It receives voice data from the client, transcribes it in real-time, and returns the text to the client.

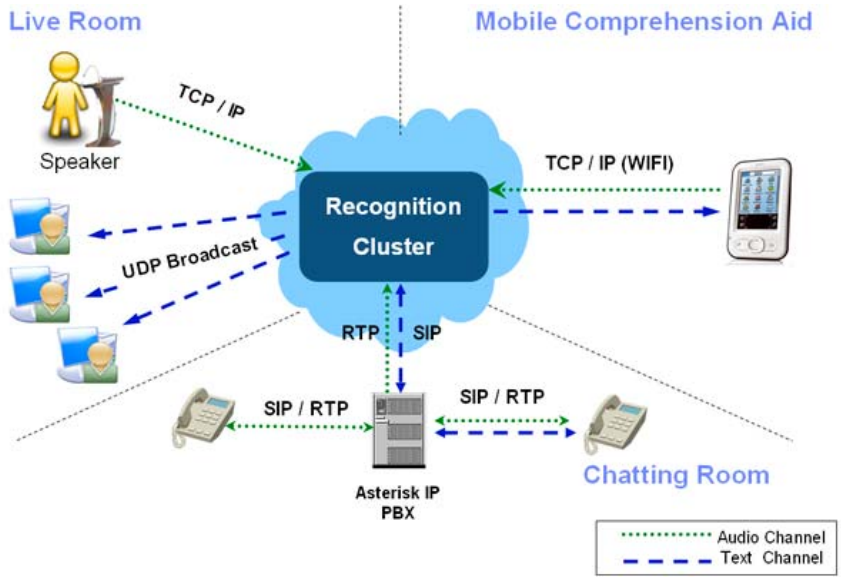

Figure 1. Paradigm of the real-time speech transcription system.

Three applications were developed to meet the requirements of different communication scenarios:

- *Live Room* for the one-way communication (lecture or seminar) scenario, where there is a primary speaker and a number of listeners that can be located in different places. The client of the primary speaker first sets up a TCP/IP connection with the server, and then begins to send the voice data of the speaker to the server. The real-time transcription generated by the server is broadcast to all listeners' client via UDP protocol. Each listener can access the text by opening the broadcast URL in a web browser during the communication.
- *Chatting Room* for the long-distance two-way communication scenario, where there are two or more participants engaged in a discussion. The application is built on a VoIP platform. During the communication, the server accesses the voice stream from the

PBX server, and returns the real-time transcription to each VoIP client via SIP/RTP protocol. Each participant can see the transcription at the client, in the chatting window of a VoIP software or on the screen of a VoIP phone.

- *Mobile Comprehension Aid* for the face-to-face communication scenario, which can happen at any time and any place. The client can be installed in any mobile phone that supports a wireless connection. When a non-native speaker has difficulty in following the other party in a face-to-face discussion, he/she can pass the mobile phone to his/her partner who will then talk to the mobile phone to get the speech transcribed. The client will communicate to the server via TCP/IP protocol, and show real-time transcription of the speech on the screen of the mobile phone.

4. Summary

We proposed providing real-time speech transcription service as a solution to improving non-native speakers' listening comprehension. First the value of transcription was demonstrated, and then a prototype system utilizing speech recognition technology to provide transcription service was presented. In future work, the effect of variables such as recognizer performance and output latency relative to the audio channel will be studied.

References

[1] 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).

[2] 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.

[3] Imoto K., Sasajima M., Shimomori T., et al. A Multi-modal Supporting Tool for Multi-lingual Communication by Inducing Partner's Reply. In *Proc. IUI'2006*, ACM Press (2006), 330-332.

[4] 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.

[5] 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.