Classroom Note-taking System for Hearing Impaired Students using Automatic Speech Recognition Adapted to Lectures

Tatsuya Kawahara, Norihiro Katsumaru, Yuya Akita, Shinsuke Mori

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

We are developing a real-time lecture transcription system for hearing impaired students in university classrooms. The automatic speech recognition (ASR) system is adapted to individual lecture courses and lecturers, to enhance the recognition accuracy. The ASR results are selectively corrected by a human editor, through a dedicated interface, before presenting to the students. An efficient adaptation scheme of the ASR modules has been investigated in this work. The system was tested for a hearing-impaired student in a lecture course on civil engineering. Compared with the current manual note-taking scheme offered by two volunteers, the proposed system generated almost double amount of texts with one human editor.

Index Terms: hearing impaired, note-taking, automatic speech recognition, lectures, adaptation

1. Introduction

Recently, more and more hearing impaired students are admitted to colleges and universities. It is imperative that schools provide necessary means to these students so that they can study just as alike as non-handicapped students. Conventionally, possible solutions are among sign language, PC captioning, and hand-writing. Currently in colleges and universities in Japan, hand-writing is most widely-used for note-taking for hearing impaired students, because it is the easiest to learn and deploy, while some schools adopt PC keyboard-typing. [1] Note-taking is conducted by student volunteers, not professional stenographers because of the budget problem. And real-time transcription of lectures is so difficult and stressing that it is widely known that only 20-30% of utterances can be transcribed even with two volunteers engaged in turn in a lecture.

Moreover, many lectures at universities are so technical that "out-of-field" volunteers cannot catch the content or technical words, for example, engineering students cannot help medical students. Actually in our university, note-takers for lectures in higher grades even in

the undergraduate can be collected only from the senior students of the same department, but it is very difficult to get a sufficient number of volunteers for all time slots.

Therefore, the automatic speech recognition (ASR) technology is expected to provide an alternative solution to assist note-taking for hearing impaired students. There have been reported trials using an ASR system for this purpose, but two persons were engaged in turn in rephrasing (or respeaking) the lecturer's utterances to a commercial dictation software program, and then the ASR result was corrected by another person, requiring three people in total. The goal of our system is to directly transcribe lecturer's utterances, so that it can be operated by one human editor.

There have been several research projects on ASR of classroom lectures [1][2][3][4], but almost all of them are designed for efficient access to audio/video archives of lectures, not real-time transcription. For real-time transcription, or efficient decoding with high accuracy, we should prepare compact acoustic and language models matched to lectures. Thus, we adopt a scheme to adapt acoustic and language models to every course and lecturer offline. Normally, one lecture course lasts a dozen of weeks, and it is often the case that the course is taught by the same lecturer for many years. Therefore, it is possible to prepare the ASR model for each lecturer and course. We can even assume that audio recording with a manual transcript of one or two hours is available. The issue of preparing transcripts for adaptation in an efficient manner is investigated in Section 3.

The output of the ASR system is selectively corrected by a human editor, before presenting to the students in need. We have implemented dedicated interfaces for making the post-editing process efficient, which will be described in Section 4. The system was tested for a hearing impaired student in real lectures in our university. Detailed analyses of the trials are reported in Section 5.

2. System Overview

An overview of the proposed system is depicted in Figure 1. The lecturer's speech is input to a microphone. Usually in classrooms, lecturers want to move around with their hands free, so we adopt a wireless pin micro-

---

[1] Real-time typing is not easy for Japanese language, because we need to convert kana (phonetic) symbols to kanji (Chinese) characters.
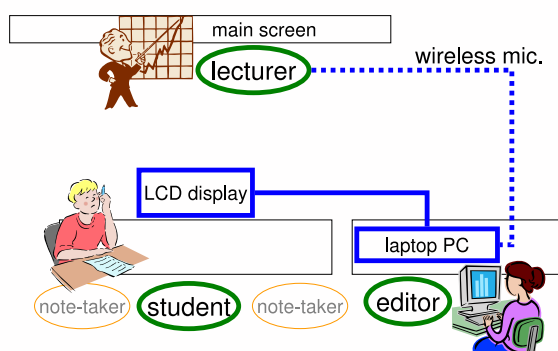
Figure 1: System overview (human note-takers are present during the trial phase)

phone. Speech is transmitted to a receiver in the same room, and then fed into a PC the system resides.

The input speech is processed by the ASR system which is adapted to the lecture. Specifically, we adapt the acoustic model to the speaker, and the language model to the content of the lecture as well as to the speaker. The system is based on our open-source ASR engine Julius [5] [2], which anyone can easily install for free.

The ASR output is generated by the utterance unit, segmented by a long pause, and given to the post-editing system, with which selection and correction is conducted. Final presentation is done through another free software program IPtalk [3], which is widely used in Japan as a PC captioning program for hearing impaired people.

The ASR system and post-editing system can operate with socket connection in a single laptop PC. An LCD screen for presentation of the text can be connected to the same PC. Thus, the whole system consists of a laptop PC and an LCD screen. This system can be extended to two PCs, which share the transcript to be corrected by two persons in cooperation. The scheme is commonly adopted in the PC captioning using IPtalk.

3. Adaptation of ASR Models

We first investigate effective methods of ASR model adaptation. If we can access to digital text media of the textbooks or slides used in the lecture, we can exploit it for adaptation of the language model. We proposed several adaptation methods based on PLSA and relevant Web texts [6]. They are effective, especially in improving keyword detection accuracy, but are limited by nature because of the small size of relevant texts.

A more effective but costly method is to use speech data given by the same lecturers, for example, in previ-

ous lectures. It is possible to record up to dozens of hours of lectures, and supervised adaptation using them would drastically improve the ASR accuracy as shown in [2]. However, it is not a practical assumption, at least in terms of budget, to prepare manual transcripts of such a large amount of data for every course. And we need to operate the system even before a large amount of data is collected. In this experiment, therefore, we prepare manual transcripts of two previous lectures (three hours in total), and compare several adaptation methods of unsupervised and semi-supervised fashions.

In the system described in Section 2, the ASR results are selected and corrected by a human editor. As the corrected text is usually cleaned and shortened, it is not a faithful transcript of the utterance, thus not suitable for model adaptation. Moreover, many ASR results are discarded if they contain too many errors or do not contain meaningful content. Still, we can make use of the information of human selection and correction. Here we assume that the human editor selects ASR results based on the word accuracy, and thus we can use only selected texts for the model adaptation.

The result of language model adaptation is summarized in Table 1, when we fix the acoustic model adapted with the same previous lectures. The baseline acoustic model (triphone HMMs of 192K Gaussians) and language model (trigram of 50K-vocabulary words) were trained with the Corpus of Spontaneous Japanese (CSJ) [7]. The table shows the standard word accuracy and the keyword detection accuracy (F-measure). The set of keywords are defined as content words that appeared in the slide text used in the lecture. The test lecture of 90 minutes was on material science for civil engineering by one lecturer.

The naive unsupervised adaptation which uses all ASR results as they are (*ASR all*) realizes a modest improvement in accuracy. We can conduct another way of unsupervised adaptation by filtering the ASR results based on the confidence measure (CM) of the recognizer (*ASR select-CM*). It is compared with supervised filtering based on the oracle word accuracy (*ASR select-oracle*). The selection was done by the utterance unit. The threshold for the selection was tuned a posteriori, and the best performance was gained by selecting 50% of the ASR results in both cases; if we select more (even by oracle), erroneous texts caused an adverse effect. As shown in Table 1, the CM-based selection is not effective at all, while the oracle selection brought improvement in keyword detection.

However, the improvement is not so significant as the case using manual transcripts or supervised adaptation (*manual*), which improved the standard word accuracy by 7.6% and keyword detection rate by 20% absolute.

The cost of manual transcription can be reduced by using the information of human editor's selection. If

Table 1: Effect of model adaptation on ASR accuracy

method	word acc.	keyword
baseline	56.3%	63.1%
ASR all	58.6%	68.0%
ASR select-CM	57.8%	68.1%
ASR select-oracle	58.2%	73.1%
manual	63.9%	83.1%
ASR select-oracle + manual	63.0%	80.8%

we use 50% of the ASR results selected by oracle (human editor), and manually transcribe the other half (*ASR select-oracle + manual*), we can achieve almost comparable performance to the supervised adaptation with the entire manual transcripts. Note again that the selection is manual, but naturally included in the system operation.

4. Post-editing Interface

We have designed and implemented post-editing interface programs. One is based on GUI, shown in Figure 2, offering several options for correction: selecting from multiple word candidates, correcting word by word, or correcting a whole sentence. This interface is functional, but human editors need some training before getting accustomed to it.

The other is based on a sequence of line editors to correct sentence by sentence. It is simple and included as one of the functions of IPtalk, so it can be used without extra training.

5. Experimental Evaluation

We conducted trials of the system in real lectures of a course on material science for civil engineering. A hearing impaired student attendeding this course was assisted by two human note-takers, who sat down next to the student and wrote down the content of the lecture by hand. In the trials, we set up our system nearby and placed an LCD screen in fron of the student. Thus, he could see either the screen or the paper notes. The experimental scene is just as described in Figure 1.

In this section, we report detailed analyses made on a lecture of 90 minutes. The test lecture is different from the lecture used in Section 3, but given by the same lecturer during the same course. We used the acoustic and language models adapted by using previous lectures of three hours. The GUI-based interface was used for post-editing in this experiment.

The amount of texts transcribed and shown to the student is compared in Table 2. We compute the ratio of the number of output words against that of all uttered words. [4] It is shown that the amount of the texts made by

---

[4] Many of the uttered words are redundant or non-sense, so even the perfect note-taking would make much smaller than 100%.
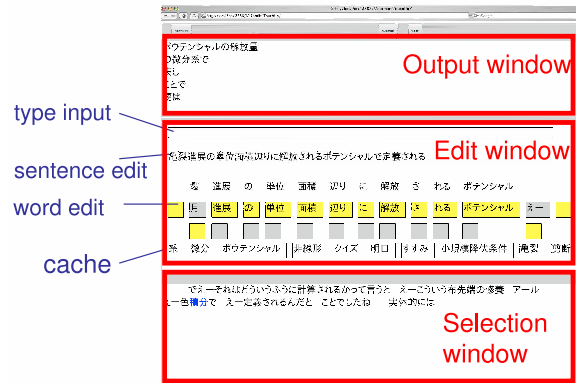

Figure 2: GUI for post-editing

our system and one human editor is significantly larger than the texts transcribed by two persons in cooperation. The hand-written texts constituted less than 20% of the original utterances, as often pointed out in Japanese. The ASR-based system restored 30-40% of the content, which might be comparable to well-trained type-writers. But most of the people cannot type-in for so long without break, so usually two or four persons are necessary for a lecture. On the other hand, our system could be operated by a single person for 90 minutes. This is a clear advantage of the system. The result also suggests that if the system was operated by two persons in cooperation, most of the content could be presented.

Figure 3 shows distributions of word accuracy for all input sentences and presented sentences. It is clear that sentences with lower accuracy (less than 40%) are mostly not used, but it is also observed that many sentences with very high accuracy are not presented; for example, more than a half of sentences with 90-100% correct are discarded. Many of them are redundant or irrelevant in terms of content, but the result suggests that the editor could not process all inputs, and the two-person scheme would improve very much.

We also investigate the latency time caused by the post-editing, as the ASR itself was performed almost in real time. The GUI-based system records the exact time (1) when it receives the ASR output, (2) when the human editor selects utterances for correction, and (3) when the editor finishes the correction and outputs the text for presentation. The average time for selecting texts ((2)-(1)) was 4.07 seconds and it is not so correlated with the ASR accuracy. The average time for correcting texts ((3)-(2)) was 4.84 seconds, and distributions of the correction time and the ASR accuracy is shown in Figure 4. We can see the general tendency that more time is needed when the ASR accuracy is lower. We can expect prompt output when the word accuracy exceeds 80%.

In order to get a feedback by the actual user, we asked

Table 2: Comparison of amount of presented texts

	words	keywords
hand-writing note-takers	16.4%	16.4%
ASR-based system	29.3%	42.9%

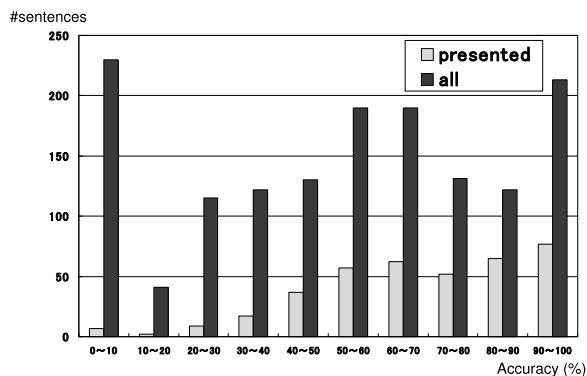Figure 3: Distribution of presented sentences and all sentences in terms of ASR accuracy

Figure 4: Distribution of correction time in terms of ASR accuracy

the university staff interview the student after the lectures. His overall impression was the system generated significantly more content than the current note-takers, but he would like much faster output though the delay by our system is not so bad as the current manual scheme. Therefore, it is foremost important to improve the ASR accuracy. [5]

6. Conclusions

In this paper, we report our effort on developing an ASR-based system for assisting hearing impaired students in classrooms. For classroom lectures, it is practical to adapt the acoustic and language models for individual lecturers, and we address an efficient scheme for the adaptation.

We made a prototype system based on the ASR and the post-editing interface, and conducted field-trials in our university. The system operated by one human editor generated twice more amount of output texts than the current manual note-taking scheme by two persons. The detailed analyses suggest that much more content could be presented if operated by two editors, and that the delay caused by the post-editing is correlated with the ASR accuracy. Thus, it is important to improve the accuracy, to the level of 80%, for improving both the latency and amount of output texts.

Currently, we have only one hearing impaired student in our university, so it is not easy to make large-scale experiments. Still, this initial result is encouraging and pro-

vides useful information for further research. In order to conduct more trials in other sites, it is crucial to realize the efficient model adaptation scheme and easy system deployment.

8. References

[1] A.Park, T.Hazen, and J.Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *Proc. IEEE-ICASSP*, volume 1, pages 497–500, 2005.

[2] J.Glass, T.J. Hazen, S.Cyphers, I.Malioutov, D.Huynh, and R.Barzilay. Recent progress in the MIT spoken lecture processing project. In *Proc. INTERSPEECH*, pages 2553–2556, 2007.

[3] I.Trancoso, R.Nunes, L.Neves, C.Viana, H.Moniz, D.Caseiro, and A.I.Mata. Recognition of classroom lectures in European Portuguese. In *Proc. INTERSPEECH*, pages 281–284, 2006.

[4] H.Yamazaki, K.Iwano, K.Shinoda, S.Furui, and H.Yokota. Dynamic language model adaptation using presentation slides for lecture speech recognition. In *Proc. INTERSPEECH*, pages 2349–2352, 2007.

[5] A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In *Proc. APSIPA ASC*, pages 131–137, 2009.

[6] T.Kawahara, Y.Nemoto, and Y.Akita. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proc. IEEE-ICASSP*, pages 4929–4932, 2008.

[7] Sadaoki Furui and Tatsuya Kawahara. Transcription and distillation of spontaneous speech. In J.Benesty, M.M.Sondhi, and Y.Huang, editors, *Springer Handbook on Speech Processing and Speech Communication*, pages 627–651. Springer, 2008.

[8] H.Nanjo and T.Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, 12(4):391–400, 2004.

---

[5]We made another experiment on an oral presentation with a proceedings paper. Once the language model was adapted with the proceedings paper, the ASR accuracy was almost 90% and texts of almost all utterances 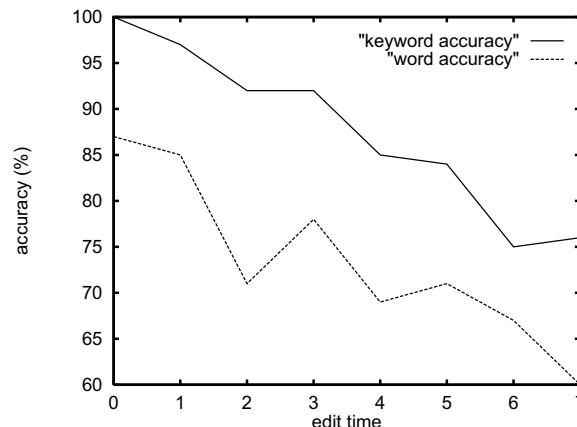were output with the simple post-editing interface.