Evaluation of the Usefulness of Spoken Term Detection in an Electronic Note-Taking Support System

Chifuyu Yonekura*, Yuto Furuya*, Satoshi Natori*, Hiromitsu Nishizaki† and Yoshihiro Sekiguchi†
* Department of Education, Interdisciplinary Graduate School of Medicine and Engineering,
† Department of Research, Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi, Kofu-shi, Yamanashi, Japan
E-mail: {yonekura,furuya,natori,nisizaki,sekiguti}@alps-lab.org Tel/Fax: +81-55-220-8361

***Abstract*—The usefulness of a spoken term detection (STD) technique in an electronic note-taking support system is assessed through a subjective evaluation experiment. In this experiment, while listening to a lecture, subjects recorded electronic notes using the system. They answered questions related to the lecture while browsing the recorded notes. The response time required to correctly answer the questions was measured. When the subjects browsed the notes, half of them used the STD technique and half did not. The experimental results indicate that the subjects who used the STD technique answered all questions faster than those who did not use the STD technique. This indicates that the STD technique worked well in the electronic note-taking system.**

I. Introduction

The primary goal of spoken term detection (STD), which is a spoken document retrieval technique, is to precisely indicate the locations (utterances) when a queried term is uttered in a large speech corpus. There have been a significant number of STD studies [1], [2], and their achievements are well reported. However, most STD studies have focused on improvement in term detection performance. In contrast, few STD studies have focused on its usefulness in a speech search system that has been put to practical use in real environments.

STD techniques may be useful in a variety of applications. For example, they can be used to search target statements from conference minute speeches. However, although there are some application areas for STD techniques, the overall usefulness of STD has not been evaluated in information systems that are of practical use in real environments.

This study evaluated the usefulness of an STD technique using an electronic note-taking support system we developed [3]. We used a previously reported STD technique [2]. The technique was installed to the note-taking support system. A user of the note-taking support system can write phrases (or figures) electronically while listening to a target speech. At the same time, the system records and stores the entire speech. Therefore, the user can review notes while listening to the recorded speech. It may also be useful to play back a speech beginning at a time specified by the time location of a note associated with a word the user wishes to focus on. The STD technique is used to indicate the location of the specified term, and it may also be useful for browsing notes associated with a speech.

This study reports the findings of a subjective experiment for assessing the usefulness of STD. In the experiment, subjects responded to questions related to a recorded speech while referring to recorded notes and listening to the speech. The subjects' response times for each correct answer were measured. Half of the subjects browsed their notes using the STD technique; the others did not use the STD technique.

The experimental results show that the subjects who used the STD technique answered all questions faster than those who did not use the STD technique. These results indicate that the STD technique works well for browsing the electronic note-taking support system.

II. STD engine

In this study, we used an STD engine employing a subword-based confusion network (CN) [2]. The engine also uses a phoneme transition network (PTN)-formed index derived from multiple automatic speech recognition (ASR) system 1-best hypothesis and an edit distance-based dynamic time warping framework to detect a queried term.

PTN-based indexing originates from the concept of CN generated from an ASR system. CN-based indexing for STD is a powerful indexing method because CNs have abundant information compared with the 1-best output from the same ASR system. Furthermore, many candidates are obtained by one or more ASR systems having different language models (LMs) and acoustic models (AMs).

The note-taking support system employs 10 types of ASR systems, with the same decoder used for all types. Two types of AMs (triphone- and syllable-based hidden Markov models) and five types of LMs (word- and subword-based) were prepared. The multiple ASR systems can generate a PTN-formed index by combining subword (phoneme) sequences from the output of the ASR systems into a single CN.

The STD performance of our STD engine demonstrated the best performance [4] among all the STD engines at the NTCIR-9 SpokenDoc STD sub-task [1].

III. Note-taking support system

The electronic note-taking support system has been developed to help students take notes during a classroom lecture.

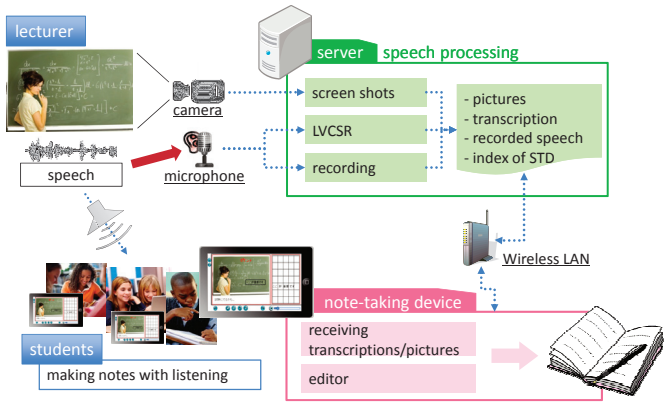

Fig. 1. Outline of the electronic note-taking support system.

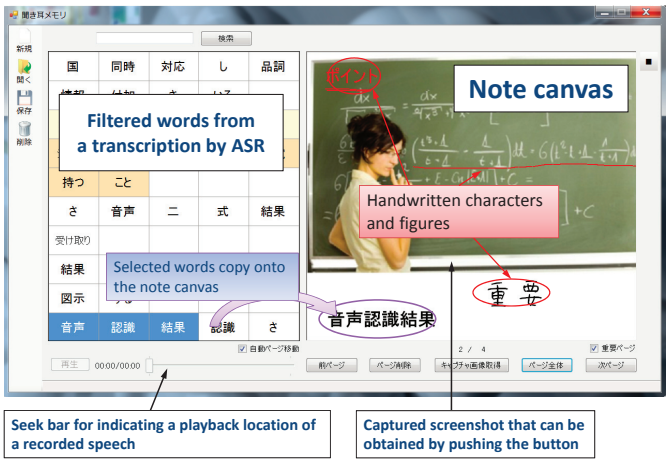

Fig. 2. Screenshot of the note-taking device.

The basic features of the system are recording a lecture and recording notes with ASR support.

A. System outline

Figures 1 and 2 show the processing flow of our note-taking support system and a screenshot of the note-taking device, respectively. The newest version of the system, which has been upgraded from a previously reported version [3], was used in this study. As shown in Figure 1, a speech is automatically recorded by a server, which also captures screenshot images of a projector screen (or a blackboard) at 3-s intervals. When the speech is stored on the server, it is transcribed by an ASR system. The transcription is then transferred to a note-taking device. At this time, the speech is also converted to a PTN-formed index for STD using the 10 ASR systems.

The word sequences generated by the ASR system are filtered by rules on the note-taking device and are displayed by the note-taking device. The user can record a note by simply touching or tracing the words on the screen. Therefore, handwriting and keyboard operations are unnecessary. The entire speech is recorded on the server, and the locations of recognized words in the speech are identified using word-to-speech alignment information produced by ASR. After the note-taking work is completed, the recorded speech is transferred to the device. Therefore, users can easily play the speech back from any chosen point. It is not necessary to listen to the entire speech when the user checks a note.

B. Speech interface

Our note-taking support system also has a speech interface. This speech interface captures the speech, records notes, and performs speech recognition. The recognized words are displayed on the screen. The user can record a note by simply touching or tracing the relevant words with their finger. Therefore, this reduces the burden on the user when recording notes while listening to a speech.

However, this system depends on the speech recognition. If the words that the user wants to associate a note with are not recognized correctly, the note cannot be recorded. Speech recognition errors are unavoidable. Therefore, it is undesirable to completely trust the ASR technology.

The key concept of the system is to use ASR as an accessory function. The ASR system is designed to help users record notes; however, it must disturb their work as little as possible. If the words that the user wants to associate a note with are not recognized correctly, the user does not need to touch or trace the words on the screen. Rather, the user can record a note using a keyboard or an electronic pen. This concept of using ASR as an accessory function differs from other systems with speech interfaces. The usability of the systems depends entirely on the output of an ASR system, and the system's usability becomes worse if ASR performs poorly.

Speech is recorded simultaneously as ASR operates. Therefore, users can listen to the speech many times if needed. Each annotated word has information that describes where the word is located in the speech. Therefore, users can easily play the speech beginning at any specified word.

C. Keyboard and handwritten input

As described in Section III-B, a user can record a note using a (hardware or software) keyboard or an electronic pen in addition to ASR. If words are faultily transcribed by ASR, the user may input the words using a keyboard or by writing.

As shown in Figure 2, a user can draw graphics, such as underlines, on the screen. A user can also write words and circle finger-touched or handwritten words for emphasis.

In this study, all the handwritten and keyboard-input words are called objects. If an object is drawn on the system's screen, it is correlated with the relative time from the beginning of the recorded speech. Therefore, if a user touches an object, they can listen to the speech starting from the object-specified time. In the subjective experiment described in Section IV, the subjects were permitted to use this function.

D. Note reference

A user can see a recorded note while listening to the recorded speech. As mentioned in Sections III-B and III-C, a user can also play the speech beginning at the time specified by the location time of the word or object.

In addition to the notes recorded by the user, all the recognized words from the speech are stored in the system. This enables the user to search for a specified word from the

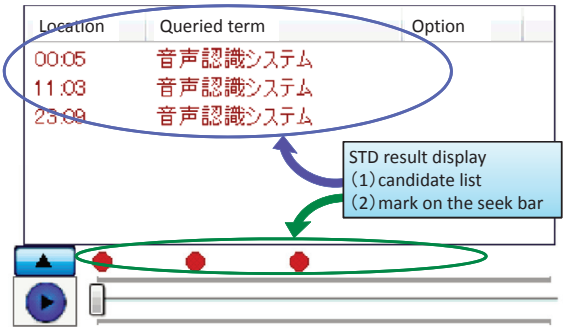

Fig. 3. STD search results interface.

recorded speech. When the user enters a word to search in the search window, the word's locations are identified if it is recognized correctly.

However, we can never search for words transcribed incorrectly. Therefore, we installed the STD technique [2], which is sufficiently robust to handle speech recognition errors.

E. STD search interface

When a user searches a queried term using the STD engine, the search results are shown on the speech playback seek bar. Figure 3 shows an example of the interface for outputting STD search results. The search results are shown in list form. Furthermore, the expected locations of the queried term, as determined by the STD engine, are indicated on the seek bar (small circle).

IV. Subjective Experiment

We performed a subjective experiment to evaluate the usefulness of STD. In the experiment, subjects took notes using the note-taking support system while listening to a lecture. Then, they answered questions related to the lecture while referring to their recorded notes and the recorded lecture. The response time required for a subject to answer all questions correctly was measured. In addition, all subjects were asked to reply to a questionnaire.

A. Experimental set up

Ten subjects, who were all university students, took notes using the system while listening to a lecture. One month after completion of note taking, the subjects answered eight questions regarding the content of the lecture. They were permitted to listen to the recorded speech when they answered the questions. At this time, half the subjects were permitted to use the STD search while the others were not. We measured the time required for a subject to answer all the questions correctly for the lecture. In this study, the response time is used as a measure of the usefulness of STD.

Table I shows the experimental conditions. AM and LM shown in the table were used for speech recognition but not for STD indexing. "Corr." and "Acc." indicate the correct word rate and word accuracy rate, respectively. "OOV" denotes the out-of-vocabulary rate for the ASR dictionary depending on LM. As shown in Table I, it was very difficult to transcribe the lecture speech using the ASR system from an ASR task point of view because Corr. and Acc. were very low.

TABLE I
Conditions of the note-taking experiment

Subjects	10 university students
Lecture name	human-machine interface
Duration	70 minitues
Recoding device	lapel microphone
ASR system	Julius rev. 4.1.3 [5]
AM	triphone-based HMM trained from CSJ [6]
LM	word-trigram (trained from a text-book and WEB documents)
Corr./Acc.	26% / 9%
Dicionary size	642 words
OOV	45%

TABLE II
STD performance

STD performance	Recall:61%, Precision: 10%
The number of STD trial	108 times (22 times / person)
Kinds of queried term	49 terms
Search time	10 s/trial

For STD indexing, we prepared two types of AMs and five types of LMs. Ten ASR systems output the transcriptions of the speech, and the PTN-formed index was created from these transcriptions. The details of the AMs and LMs are described in a previous report [2].

B. Experimental result

Table II shows how many times the subjects (half of the subjects) attempted to use the STD search, its term detection performance, and search processing time. The recall and precision rates, which are popular evaluation measures for STD [1], were 61% and 10%, respectively. These values are higher than expected because the ASR performance was very low. In addition, it was very hard to detect two queried words that consist of less than five phonemes. By removing the queried terms from the STD performance calculation, the recall and precision rates became 67% and 20%, respectively.

Table III shows the average response times and standard deviations for all questions for each subject group. "group w/o STD" indicates a group that did not use the STD search and "group w/ STD" indicates a group that used the STD search. As shown in Table III, "group w/ STD" answered the questions faster than "group w/o STD"; however, there was no significant difference between the average response times. Furthermore, the standard deviation of "group w/ STD" is less than that of the other group. This indicates that there were few differences between individuals. Table IV shows the average response time required to answer each question for each group. As shown in Table IV, for questions #2, #5, and #7 with symbol "*", the response times for "group w/ STD" were clearly shorter. In addition, there were no statistical differences between the response times of either group for the other questions. This indicates that the STD technique was useful for effectively searching target terms from the recorded lecture speech.

C. Questionnaire result

Table V shows the questionnaire results related to the STD search. The subjects belonging to "group w/ STD" answered the questionnaire, which evaluated the following items: (1)

TABLE III
AVERAGE RESPONSE TIMES REQUIRED BY EACH GROUP TO ANSWER QUESTIONS CORRECTLY [MM'SS"]

	group w/o STD	group w/ STD
Average	40'58"	35'25"
Standard dev.	14'34"	6'17"

TABLE IV
AVERAGE RESPONSE TIMES REQUIRED BY EACH GROUP TO ANSWER EACH QUESTION CORRECTLY [MM'SS"]

Question	group w/o STD	group w/ STD
#1	2'27"	2'46"
#2	6'33"	* 1'28"
#3	5'11"	6'18"
#4	2'43"	3'49"
#5	14'33"	* 11'55"
#6	3'02"	4'03"
#7	3'55"	* 2'05"
#8	2'34"	3'00"

search speed and (2) STD necessity. Each item was evaluated on a five-point scale (from 1 to 5), with greater numbers indicating more positive evaluation.

Comments from "group w/o STD" are as follows:

- There are a lot of ASR errors.
- When a target term is missed by ASR, I had to listen to the entire speech. In this case, it was tough.

On the other hand, comments from "group w/ STD" are as follows:

- The STD search was useful for finding the target location in the speech when the target term was missed by ASR.
- The seek bar with symbols from the STD engine helped me search for the location of information required to answer questions quickly.
- STD search time (turnaround time) was very slow.

D. Discussion

The average response time of "group w/ STD" was shorter than that of "group w/o STD." Therefore, with a 61% recall rate, the STD technique was useful for browsing the notes for a lecture, for which speech recognition was difficult (Corr. and Acc. were 26% and 9%, respectively).

Furthermore, as shown in Table III, the standard deviation of "group w/o STD" is high. Therefore, there is no statistically significant difference between the response times of the two groups. It is evident that there are a few differences between individual response times for "group w/ STD" because the standard deviation of this group is less than that of the other group. This indicates that the STD search enables an efficient reference work for an electronic note regardless of individuals. The average rating from the questionnaire related to the usefulness of STD was 4.2. Thus, the subjects indicated that the STD technique was useful.

On the other hand, there were no improvements for the response times for questions #1, #3, #4, #6, and #8. This is because the STD search processing time was slow (10 s/trial). If the processing speed can be improved, the usefulness will also increase.

From the comments provided by the subjects, the STD search was used for two purposes. One was to search a term that was missed by ASR when they recorded a note and the other purpose was to indicate the range of speech they want to listen to. The latter purpose is very interesting. Visual display of the STD result on the seek bar was a very convenient method to narrow the range of speech. The STD performance for a speech recorded in a real environment (lower S/N ratio) will be worse than an ideal environment (higher S/N ratio). However, improving the STD interface can compensate for degradation of STD. This is an interesting finding.

TABLE V
QUESTIONNAIRE RESULTS RELATED TO STD (AVERAGES FOR SUBJECTS)

Question	(1) search speed	(2) necessity
Ratings	1.8	4.2

Finally, STD search speed is a very important factor in practical application of an STD technique. The processing speed of our STD engine was slow. However, we intend to reduce the processing time to 3 s by improving the STD implementation in the note-taking device.

V. CONCLUSION

This study introduced an electronic note-taking support system and investigated the usefulness of an STD technique.

In the subjective experiment, the subjects took notes using the note-taking support system while listening to a lecture. One month after finishing the work, they answered questions related to the lecture while browsing the recorded notes. The response time required to answer all questions correctly was measured. In addition, the subjects replied to a questionnaire regarding the STD search. The experimental results show that the response times when using the STD search were shorter than those without the STD search for specific questions, although speech recognition of the lecture was difficult. Furthermore, we found that the design of an STD search interface is important to facilitate sufficiently efficient STD.

In future, we plan to improve the STD engine to reduce search time.

REFERENCES

[1] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the IR for spoken documents task in NTCIR-9 workshop," in *Proceedings of NTCIR-9 Workshop Meeting*, 2011, pp. 223–235.

[2] S. Natori, Y. Furuya, H. Nishizaki, and Y. Sekiguchi, "Spoken Term Detection Using Phoneme Transition Network from Multiple Speech Recognizers' Outputs," *Journal of Information Processing*, Vol.21, No.2, 2013, pp. 176–185.

[3] K. Ota, H. Nishizaki, and Y. Sekiguchi, "Development of Note-Taking Support System with Speech Interface," in *Proceedings of APSIPA ASC 2012*, 2012.

[4] H. Nishizaki, Y. Furuya, S. Natori and Y. Sekiguchi, "Spoken Term Detection Using Multiple Speech Recognizers' Outputs at NTCIR-9 SpokenDoc STD subtask," in *Proceedings NTCIR-9 Workshop Meeting*, 2011, pp. 236–241.

[5] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009)*, 2009, pp. 131–137.

[6] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*. ISCA, 2003, pp. 7–12.