WESPAC IX 2006
The 9th Western Pacific Acoustics Conference
Seoul, Korea, June 26-28, 2006

COMPUTER ASSISTED SPEECH TRANSCRIPTION SYSTEM FOR EFFICIENT SPEECH ARCHIVE

Hiroaki Nanjo[†], Yuya Akita[‡] and Tatsuya Kawahara[‡]

†*Faculty of Science and Technology, Ryukoku University*
Seta, Otsu 520-2194, Japan
nanjo@ryukoku-u.jp

‡*Academic Center for Computing and Media Studies, Kyoto University*
Sakyo-ku, Kyoto 606-8501, Japan
yuya@media.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

ABSTRACT

This paper addresses computer assisted speech transcription (CAST) system for making archives such as meeting minutes and lecture notes. For such system, automatic speech recognition (ASR) is promising, but ASR errors are inevitable. Therefore, it is significant to design a good interface with which we can correct errors easily. Moreover, to make a better system, we should know what kind of recognition error is significant and how much accuracy of ASR we need in correcting errors.

From these points of view, we design a CAST system with several correction interfaces; 1) pointing device for selection from confusion pairs, 2) microphone for re-speaking, and 3) keyboard. With some subject experiments, we confirmed that the system could reduce a transcription time by about half at the best case. We also found that 75% or more ASR accuracy should be achieved for users to feel ASR system is convenience in correction task.

KEYWORDS: computer assisted speech transcription, spontaneous speech recognition, error correction interface, confusion network

INTRODUCTION

According to the progress of automatic speech recognition (ASR) of human-to-human spontaneous speech, which would make possible the automatic transcription or translation of lectures and meetings, ASR accuracy of 70% to 80% has been achieved [1][2]. These figures are based on the 1-best hypothesis which is finally output by ASR decoder. In contrast, ASR accuracy based on a wordgraph, which is used to reduce a search space in a decoder, exceeds 90% even for a spontaneous speech [3]. We also have confirmed the same tendency according to our preliminary experiments. Although we can achieve quite accurate wordgraph, studies for the effective use of it are not sufficient. Thus, in this paper, we study an efficient ASR applications taking advantages of such an accurate wordgraph. Specifically, we design a computer assisted speech transcription (CAST) system which helps us to make transcriptions of spontaneous speech such as meeting minutes and lecture notes. The system should have good interfaces with which we can easily correct ASR errors, which are inevitable. In this paper, we demonstrate the effectiveness of our CAST system according to the subject experiments, then we show one goal of spontaneous speech recognition, that is, how much accuracy of ASR we need is discussed.

COMPUTER ASSISTED SPEECH TRANSCRIPTION SYSTEM

Overview. In order to make transcriptions using ASR systems, a careful error check by human is necessary because ASR errors are essentially inevitable. Therefore, it is significant to design a good interface with which we can correct errors easily. Providing several correction methods (interfaces) would be more helpful.

According these concepts, our computer assisted speech transcription (CAST) system is designed. Figure 1 shows the system overview. At the upper left side in Figure 1, ASR results for each utterance are listed. At the bottom area, a MAP decoding result (word sequence), which a user selects from the utterance list, and its competitive candidates are displayed. Users can correct ASR errors by selecting correct words from the competitive candidates. Users can listen to a recorded audio material (utterance) of a transcription target by clicking the “Play speech” button. This correction interface is similar to the “Speech Repair system” [3].

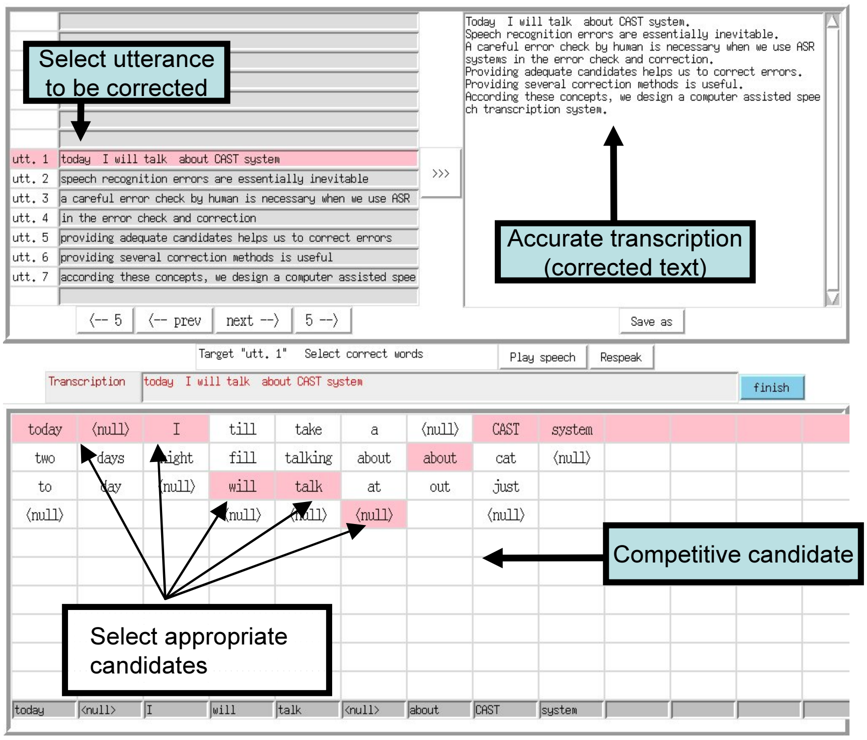
Our CAST system also provides other correction interfaces; 1) traditional correction (keyboard correction) interface and 2) speech correction interface. The former correction interface is expected to be used when there are no appropriate correction candidates in the competitive candidate area. The latter correction interface is expected to be used when there are a lot of ASR errors which causes large increase of the correction cost with a pointing device or keyboard. In respeak based correction, users speak (respeak) what the original speaker said excluding fillers and repairs, and then selects competitive candidates, which are generated by ASR decoder incorporated to the CAST system.

Specification of Speech Recognition System. The specification of ASR system, which is incorporated to the CAST system, is described. Here, we use an ASR system for meetings of the National Congress (Diet) of Japan, which we have been investigating [2].

For a language model, we collected the four-year minutes of the National Congress from the 145th ordinary session in 1999 to the 158th extraordinary session in 2003. All meetings including plenary sessions, committees and public hearing are used. The text size is 86.7M words. Text of the minutes is basically faithful record of utterances. However, it does not contain spoken expressions such as fillers, disfluencies and colloquial expressions, since these were modified or removed for documentation. To cover these expressions, we also prepared true transcription of some meetings, whose size is 353K. A back-off trigram language model is then constructed by synthesizing these two corpora. The vocabulary size of the model is 52,093.

A pronunciation lexicon is automatically generated in order to deal with pronunciation variations. In spontaneous speech, multiple pronunciations are observed for a linguistically identical word and the phenomenon degrades ASR performance. In this paper, we adopt our applicable lexicon generation method that is based on a probabilistic mapping of phone sequences using a large-scale spontaneous speech corpus [2]. As a spontaneous speech corpus, we use the CSJ (Corpus of Spontaneous Japanese) [4] [5].

Acoustic model is based on continuous density Gaussian-mixture HMM. We use a gender-independent triphone model trained with academic presentation speech of the CSJ. Total amount of training speech data is 228 hours. The numbers of shared states and mixture components are 3,000 and 16, respectively. We performed unsupervised speaker adaptation. In order to recognize respeaked speech uttered by a CAST system user, we use a gender-independent triphone model which is trained with read speech (JNAS – newspaper reading).


Select utterance to be corrected
Accurate transcription (corrected text)
Competitive candidate
Select appropriate candidates
>>>
<-- 5
<-- prev
next -->
5 -->
Save as
Play speech
Respeak
finish

utt. 1 today I will talk about CAST system
utt. 2 speech recognition errors are essentially inevitable
utt. 3 a careful error check by human is necessary when we use ASR
utt. 4 in the error check and correction
utt. 5 providing adequate candidates helps us to correct errors
utt. 6 providing several correction methods is useful
utt. 7 according these concepts, we design a computer assisted spee

Today I will talk about CAST system.
Speech recognition errors are essentially inevitable.
A careful error check by human is necessary when we use ASR
systems in the error check and correction.
Providing adequate candidates helps us to correct errors.
Providing several correction methods is useful.
According these concepts, we design a computer assisted spee
ch transcription system.

Target "utt. 1" Select correct words

Transcription today I will talk about CAST system

today	⟨null⟩	I	till	take	a	⟨null⟩	CAST	system
two	days	might	fill	talking	about	about	cat	⟨null⟩
to	day	⟨null⟩	will	talk	at	out	just	
⟨null⟩			⟨null⟩	⟨null⟩	⟨null⟩		⟨null⟩	
today	<null>	I	will	talk	<null>	about	CAST	system

Figure 1*. Overview of computer assisted speech transcription (CAST) system*

We used recognition engine Julius 3.5[6]. Here, we adopt sequential decoding so that very long speech can be handled without prior segmentation [7].

GENERATION OF COMPETITIVE CANDIDATES

Comparison of N-best List and Wordgraph. We investigated how to generate an suitable competitive candidates from ASR results for error correction. Here, N-best list based approach and wordgraph based approach were investigated. For an evaluation, audio data of the committee held on February 14, 2003 was used. They are not included in training data of the language model. There are 23 participants in this time and the duration of speech is about 5.5 hours. The total number of words is 62,512. Word accuracy and correctness based on a MAP decoding are 80.1% and 83.6%, respectively.

Next, we described a wordgraph based approach. In this work, a wordgraph is dynamically generated during 2nd pass search of the Julius. A graph accuracy and correctness are 92.5% and 93.8%, respectively. A wordgraph density, which represents an average number of candidates per word, is 54.4. When we generate a wordgraph[1] from the N-best list, a graph density is 5.9, and a graph accuracy and correctness are 88.6% and 90.8%, respectively.

We confirmed that 1) correct candidates for half of the ASR error of a MAP decoding were included in a wordgraph, 2) dynamically generated wordgraph outperformed N-best based wordgraph, and 3) we could achieve quite accurate wordgraph even if we adopt an N-best based method. These results show that ASR system is available and helpful for transcription of spontaneous speech.

Confusion Network. Although the wordgraph or N-best list are quite accurate, it is difficult to correct errors quickly even if a wordgraph or N-best list itself is shown because the competitive candidates for each

[1] To be exact, we generate an Confusion Network which is described at the following section.

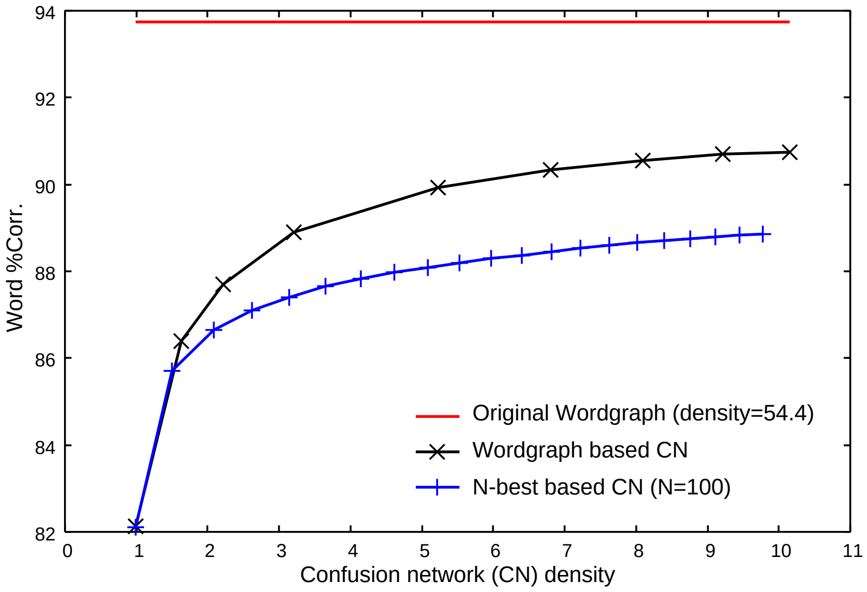


Figure 2*. Relationship between confusion network (CN) density and word correctness*

word are not clear. Thus, we adopt Confusion Network (CN) framework [8], and represent ASR results as a CN. CN is a time aligned sequence of word sets, and respective word-sets have competitive words. N-best based CN is generated by aligning N-best hypotheses to one-best hypothesis using DP matching algorithm. Wordgraph based CN is generated by clustering words based on their temporal overlap. The CN based competitive candidate representation is shown in Figure 1 at the bottom area.

Figure 2 shows word correctness of both CNs for each CN density. The wordgraph based CN outperform N-best list based CN where the density is 2 or more. CN density represents an average number of competitive candidates for each word, and CAST system users may correct errors with selecting correct words from the competitive candidates. Therefore, suitable number of candidates are required. Considering a design of human use, here we tried to generate 10 candidates for each word. When we generate wordgraph based CN whose density around 10, we achieve 91% of word correctness, which is comparable to the correctness of original wordgraph (93.8%, density: 54.4). We confirm that CN compression for the CAST system works reasonably.

EVALUATION OF CAST SYSTEM USABILITY

The CAST system was tested with 18 Japanese subjects (6 males and 12 females). Almost all of them are students of Ryukoku University, Japan.

Time for Error Correction and Transcription. Firstly, we evaluate the CAST system from a viewpoint of efficiency. We measured a time which each subject required for making transcriptions. Specifically, without the CAST system (each subject just input texts with a keyboard), the time for transcription was measured, and then the time with CAST system was measured. Each subject tests with five utterances whose durations are about 30 seconds. Their ASR accuracy is distributed among 50% to 90%. The time is measured including all processes of error correction and transcription, which is described as follows.

(1) listen to an audio material

(2) transcribe what users listen to

(3) check the text with listening to the corresponding audio material again

(4) repeat 2 to 3 stages if necessary

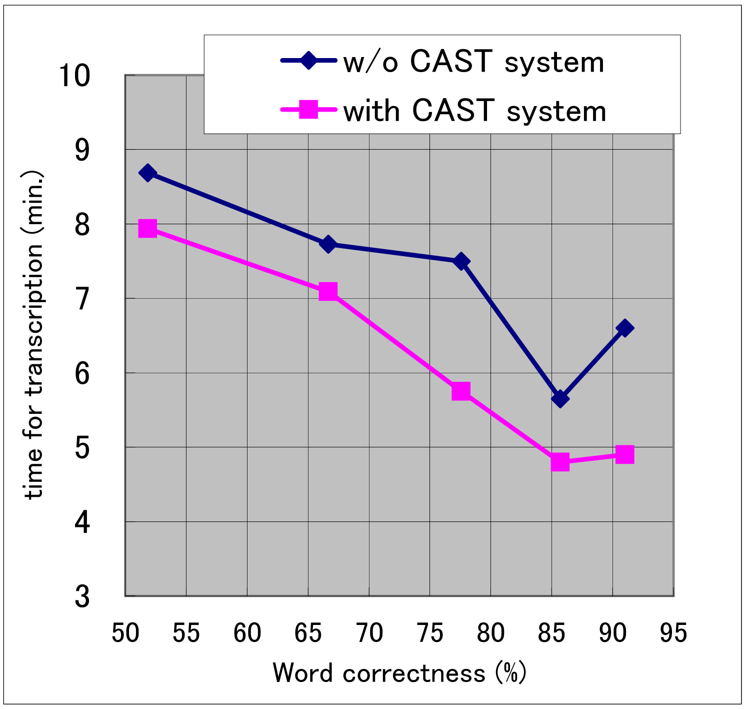


***Figure 3**. Relationship between transcription time and ASR accuracy (Word %Corr.)*

Figure 3 plots the averaged times for transcription for each utterance. With the CAST system, subjects can correct errors and make transcriptions faster than without the system. About five to eight minutes are required with the CAST system. Without the CAST system, users required extra one to two minutes. Considering above described processes of error correction and transcription, users mainly cut down the working costs at a transcription stage since a listening stage can not be omitted basically. That is, an actual transcription time can be estimated by subtracting two to four times of the audio duration from the total time which is required for error correction and transcription. For the utterance of 90% of ASR accuracy, the actual time for transcription is estimated to be three or four minutes with the CAST system. Without the system, the actual time is regarded as five or six minutes. Thus, the CAST system can reduce the time for transcription by half. These results show the effectiveness of the CAST system. Figure 3 also shows that subjects required less time to transcribe according to an ASR accuracy whether the CAST system is used or not. The result suggests that an ASR accuracy is correlated to a difficulty of listening comprehension.

Next, we discuss the system effectiveness from a viewpoint of unevenness among users. With the CAST system, ten subjects required less time, and five subjects required almost same time for transcription. Although using the CAST system, three subjects required much time. We considered one cause is that they are good at keyboard typing. Actually, two of the subjects who required much time with the CAST system belonged to the best keyboard typing group among all subjects. Another possibly cause is that users spent a time during the error correction by respeaking because they are unfamiliar to the ASR system.

These results indicates that the system is efficient and usable for most users.

Analysis of Required ASR Performance. Next, we describe how much accuracy of ASR we need in correcting errors. The analysis is performed based on the results of the subject experiments. The data size is 228 from 18 subjects. The subjects were asked for a rating of a questionnaire of “Is ASR helpful for making a transcription?” on a scale of one to seven. Relationship between the rating score and ASR accuracy (word correctness) is shown in Figure 4. Higher rating scores (5 or more) are observed where ASR accuracy is over 75%, that is, in correction task, 75% or more of ASR accuracy should be achieved

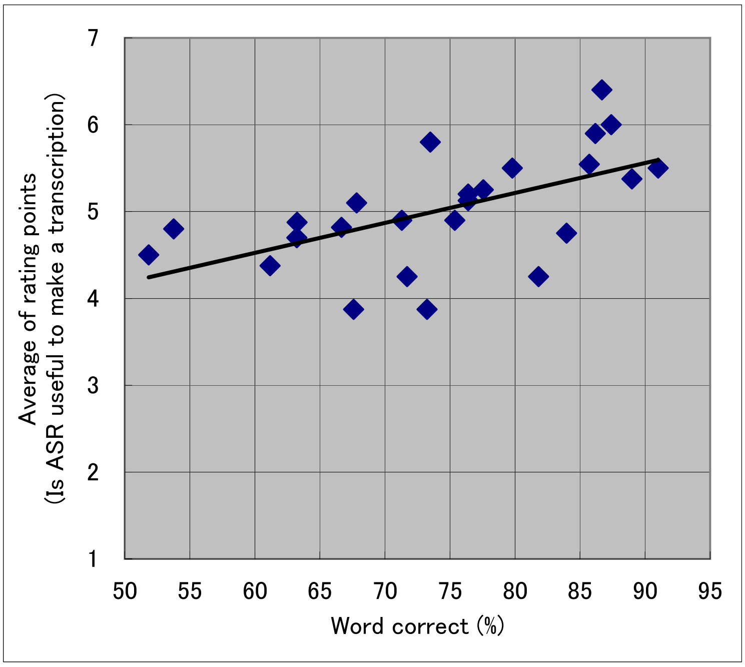


Figure 4. Relationship between rating point and ASR accuracy (Word %Corr.)

Table 1. ASR usability analysis (multiple linear regression analysis)

explanatory variable	standard partial regression coefficient	partial regression correlation
ASR performance	0.249**	0.263
Experience as a PC user	-0.189**	-0.192
Score of keyboard typing	0.143	0.121
Facility on system use	0.147*	0.144
Comfortableness on system use	0.279**	0.210

**: significant (1%), *: significant (5%)

for users to feel ASR system is convenience. The figure can be one goal of ASR of spontaneous.

Analysis of System Usability. Finally, we discuss the usability of the CAST system. Here, we also use the results of the subject experiments. The subjects are asked for a rating of some terms about the usability of the CAST system on a scale of one to seven.

We performed a factor analysis and identified five factors; 1) “fineness of system appearances”, 2) “comfortableness on the use of the system”, 3) “facility on the use of the system”, 4) “familiarity with automatic speech recognition”, and 5) “degree of exhaustion on the use of the system”.

Then, we performed multiple linear regression analysis. Criterion variable is a rating score of a questionnaire of “Is ASR helpful for making a transcription?”. Explanatory variables are keyboard typing score (strokes per minute), experience as a PC user (months), ASR accuracy, and five factors which are identified by the above described analysis. We found a multiple regression coefficient of 0.409 (significant in 1% level). Table 1 lists a standard partial regression coefficient and a partial regression correlation for each explanatory variable in a multiple regression equation. These results show that (1) people tend to find an advantage of taking ASR system according to ASR accuracy, (2) people who feel more comfortable during a system use consider an ASR to be usable, and (3) people who are less experienced in using computers have a tendency to decide that ASR is convenient.

The result that the ASR is convenient for less experienced people encourages us to make ASR system widespread.

CONCLUSION

Computer assisted speech transcription (CAST) system, which is helpful to make a transcription of spontaneous speech, is addressed. With the CAST system, users can make a transcription by correcting ASR results using suitable error correction interface which are provided by the CAST system. One of the most significant correction method is selection from competitive candidates. We investigated several generation methods of competitive candidates and confirmed the effectiveness of confusion network (CN) based generation. The CN based competitive candidates (density$=10$) include 91% of correct words and the accuracy is comparable with a wordgraph accuracy (density$=54.4$), of which the CN based competitive candidates are generated.

We also evaluated the CAST system with subject experiments. For the transcription task with some efficient correction interfaces, we found that one goal of spontaneous speech recognition is 75% of accuracy (word correctness). We also found that ASR could be useful for the people who have less experiences of computers.

Acknowledgment: The authors are grateful to the House of Representatives of Japan for providing us with speech data of meetings and partially supporting this research. The work was also partly supported by Ryukoku University Science and Technology fund.

REFERENCES

1. H.Nanjo and T.Kawahara, “Language model and speaking rate adaptation for spontaneous presentation speech recognition,” *IEEE Trans. Speech & Audio Process.*, vol. 12, no. 4, pp. 391–400, 2004.

2. Y.Akita and T.Kawahara, “Generalized statistical modeling of pronunciation variations using variable-length phone context,” in *Proc. IEEE-ICASSP*, 2005, vol. 1, pp. 689–692.

3. J.Ogata and M.Goto, “Speech repair: Quick error correction just by using selection operation for speech input interfaces,” in *Proc. EUROSPEECH*, 2005, pp. 133–136.

4. K.Maekawa, “Corpus of Spontaneous Japanese: Its design and evaluation,” in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.

5. S.Furui, K.Maekawa, and H.Isahara, “Toward the realization of spontaneous speech recognition – introduction of a Japanese priority program and preliminary results –,” in *Proc. ICSLP*, 2000, vol. 3, pp. 518–521.

6. A.Lee, T.Kawahara, and K.Shikano, “Julius – an open source real-time large vocabulary recognition engine,” in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.

7. T.Kawahara, H.Nanjo, and S.Furui, “Automatic transcription of spontaneous lecture speech,” in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

8. L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech & Language*, vol. 14, pp. 373–400, 2000.