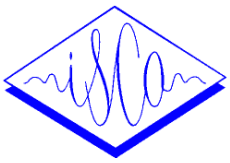

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

TOWARD THE REALIZATION OF SPONTANEOUS SPEECH RECOGNITION –– INTRODUCTION OF A JAPANESE PRIORITY PROGRAM AND PRELIMINARY RESULTS ––

Sadaoki Furui[1,2], *Kikuo Maekawa*[2], *Hitoshi Isahara*[3], *Takahiro Shinozaki*[1] *and Takashi Ohdaira*[1]

[1] Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan, furui@ cs.titech.ac.jp
[2] The National Language Research Institute
3-9-14 Nishiga'oka, Kita-ku, Tokyo, 115-8620 Japan, kikuo@kokken.go.jp
[3] Communications Research Laboratory
588-2 Iwaoka, Nishi-ku, Kobe, 651-2401 Japan, isahara@crl.go.jp

ABSTRACT

Although high-recognition accuracy can be obtained for speech in the form of reading a written text or similar by using state-of-the art speech recognition technology, the accuracy is quite poor for freely spoken spontaneous speech. From this perspective, a new national project for raising the technological level of speech recognition and understanding has recently commenced in Japan. This paper first briefly introduces the project and then reports some results of preliminary experiments which have been conducted at Tokyo Institute of Technology.

1. INTRODUCTION

Read speech or similar, such as speech reading newspapers and broadcast news utterances made by announcers, can be recognized with a higher than 90% accuracy using the present speech recognition technology. NHK broadcasting company in Japan recently introduced an online close-captioning system using speech recognition technology. However, the recognition accuracy dramatically declines for spontaneous speech. The principal reason for this is that acoustic and linguistic models used in speech recognition have been built using written language or speech reading text, while spontaneous speech and written language considerably differ both acoustically and linguistically. Broadening effectively the application of speech recognition thus crucially depends on raising the recognition performance for spontaneous speech.

From this viewpoint, a Japanese national project on spontaneous speech corpus and processing technology was initiated in 1999. This project aims to build a large-scale spontaneous speech corpus and create spontaneous speech recognition and understanding technology.

2. JAPANESE NATIONAL PROJECT ON SPONTANEOUS SPEECH CORPUS AND PROCESSING TECHNOLOGY

The Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" started in 1999 under the supervision of S. Furui. The principal organizations working together to conduct this project are National Language Research Institute under the Ministry of Education, Communication Research Laboratory under the Ministry of Posts and Telecommunications, and Tokyo Institute of Technology.

The project will be conducted over a 5-year period in pursuit of three major themes as shown in Fig. 1:

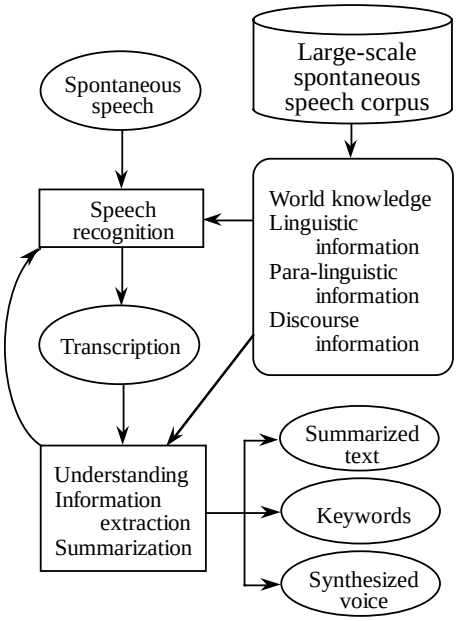

Fig. 1: Overview of the national project

1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words with a total speech length of 1000 hours. Mainly recorded will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. One-tenth of the utterances ("Core") will be tagged manually and used for constructing a morphological analysis program for automatically analyzing all of the

1000-hour utterances. The Core will also be tagged with para-linguistic information including intonation [1].

2) Acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information in speech.
3) Constructing a prototype of a spontaneous speech summarization system.

The technology created in this project is expected to be applicable to wide areas such as indexing of speech data (broadcast news, etc.) for information extraction and retrieval, transcription of lectures, preparing minutes of meetings, closed captioning, and aids for the handicapped.

Presentations at various conferences, such as the Acoustical Society of Japan (ASJ) meetings, and free presentations by voluntary subjects are recorded and transcribed in the project. Using these utterances, preliminary recognition experiments are being conducted at several universities participating in the project. At Tokyo Institute of Technology, for example, preliminary experiments have been conducted using a presentation recorded at an ASJ meeting and one person's talk excerpted from a broadcast political discussion.

3. TRANSCRIPTION OF PRESENTATION

3.1. Recognition task and experimental conditions

The first experiment was conducted using one male speaker's presentation recorded at the ASJ meeting. The content of the presentation was an overview of speech synthesis technology. The utterances were recorded using a close-talking microphone, stored on a DAT tape, and transcribed manually. The first 22 minutes and the latter six minutes were separated for training and testing, respectively.

3.2. Language modeling

Since we have no spontaneous speech corpus yet, we first employed the language model used in broadcast news transcription. This model was constructed using broadcast news texts over a 34-month period comprising 380,000 sentences. Since there is no clear definition of words in Japanese and no spacing between words in written Japanese sentences, a morphological analysis program was used to split sentences into morphemes, and the morphemes (which will be called "words" from now on) were used as units for statistical language modeling. A set of unigrams, bigrams and trigrams was calculated and smoothed utilizing the Katz's back-off smoothing technique. This language model made from broadcast news texts is referred to LM1.

In order to build a corpus more appropriate for recognizing the lecture utterances, we collected texts of transcribed lectures from the World Wide Web [2]. The texts that appeared to be relatively close to their original utterances were selected to build the corpus. Japanese texts are usually written by using the mixture of multiple sets of characters composed of Kanji, Hiragana, Katakana, Western and Japanese numerals, Western alphabets, and various symbols. Even periods and commas have multiple styles. In addition, there are two kinds of the character widths: "Zen-kaku" and "Han-kaku." Considerable freedom is permitted in using these characters to write the same word. Therefore, the characters in the texts were preprocessed automatically as well as manually to normalize the variations. The statistical features of the corpus subsequent to the preprocessing are given in Table 1.

Table 1: Corpus of lectures collected from WWW

Number of themes	43
Number of sentences	76,000
Number of words	2,000,000
Vocabulary size	46,000

This corpus was analyzed by the morphological analysis program and used to build a statistical language model we call LM2.

Spontaneous speech usually includes various filled pauses, but neither the broadcast news corpus nor the lecture corpus includes them. An effort was thus made to add filled pauses to the lecture corpus based on the statistical characteristics of the filled pauses observed in the training part of the presentation. Since it was found that the filled pauses typically occurred immediately before the sentence beginnings and before or after commas, the number of occurrences of 21 kinds of filled pauses at these positions was calculated using the training period of the presentation. According to these results, filled pauses were randomly added to the lecture corpus and a new language model was built named LM3.

3.3. Acoustic modeling

The following two acoustic models were employed:

- AM1: HMMs trained by using read speech uttered by 130 male speakers.
- AM2: HMMs trained by the maximum likelihood method using the training period of the presentation in which the AM1 was used as the initial model.

3.4. Experimental results

Table 2 presents the test-set perplexity and the out-of-vocabulary (OOV) rate for each language model. LM2 and LM3 made from the lecture corpus collected from the WWW are clearly indicated as being much more appropriate for modeling the presentation speech than LM1 made from the broadcast news texts in terms of both the test-set perplexity and the OOV rate.

Table 2: Test-set perplexity and OOV rate for the presentation speech

Language model	Test-set perplexity		OOV(%)
	Bigram	Trigram	
LM1	773.0	929.0	12.6
LM2	197.8	194.5	5.6
LM3	212.3	209.1	5.7

Table 3 shows the recognition results for the combinations of the two acoustic models and the three language models. Using LM2 reduced word error rates by 11% and 13% compared with LM1 in the cases of AM1 and AM2 respectively. Employing LM3 further reduced the error rates by 5% and 10% for AM1 and AM2 respectively. Due to the improvement of both acoustic and language models, that is using AM2 and LM3, the error rates were reduced by 28% compared with the utilization of AM1 and LM1. These results demonstrate that the models based on the proposed methods are very effective.

Table 3: Recognition results for the presentation speech

AM1

Language model	%Corr(%)	Acc(%)	%Err(%)
LM1	20.1	15.3	84.7
LM2	33.9	24.9	75.1
LM3	32.9	28.8	71.2

AM2

Language model	%Corr(%)	Acc(%)	%Err(%)
LM1	29.4	22.6	77.4
LM2	44.6	32.4	67.6
LM3	44.2	39.4	60.6

3.5. Discussion

A supplementary experiment was performed concerning the filled pause problem in the language modeling. In this experiment, the filled pauses were not separated into 21 different words but modeled as a single word having 21 pronunciations with equal probability. Experimental results showed that many words occurring at sentence beginnings and before or after commas were recognized as filled pauses, and as a result, the word error rate was increased to 63.2% with AM2. This means that the effectiveness of the filled pause modeling was halved compared with the method separately modeling the different filled pauses.

4. TRANSCRIPTION OF DISCUSSION

4.1. Recognition task and experimental conditions

The second experiment was performed using a part of a discussion enjoined by politicians which was broadcast over radio and TV simultaneously. From the 60-minute program, one politician's utterances consisting of 217 sentences were extracted and used in the experiment, from which the first 100 utterances were used for training and the latter 117 utterances were used for testing.

4.2. Language modeling incorporating out-of-vocabulary words

In addition to LM1 constructed using broadcast news texts, LM2 built from transcribed lectures collected from the WWW, and LM3 made by adding filled pauses adapted to the presentation speaker in the previous section to the transcribed lecture, the following language models were investigated to contend with the out-of-vocabulary (OOV) problem.

Although various methods for automatically detecting OOV words in the recognition utterances have been investigated, no practical way has yet been proposed. Therefore the assumption was made that most of the task-dependent OOV words, in other words "new words", are brought into the system by some method prior to recognition. With presentations, for example, a written paper corresponding to each presentation is usually printed in the proceedings and can be obtained in advance, and all the words included in the paper but not contained in the vocabulary for recognition can easily be detected.

Based on this assumption, several methods for incorporating the new words in the language models were investigated. One of the conventional methods involves approximating word n-grams of new words by using class n-grams, in which each new word is assigned to one of the pre-defined word classes. This method, however, is problematic in designing word classes. Thus, all of the OOV words in the test set were assumed to have been detected or given beforehand and the following three methods were investigated.

- LM4: Unigrams for all of the OOV words with a common empirically decided value were added to the set of unigrams of LM3 and all the n-gram values were normalized.
- LM5: In the text used for training LM3, a fixed rate (10%) of the words in the same fine morpheme class as each OOV word was randomly replaced by the OOV word, and the text was then used for building a language model.
- LM6: A single OOV class n-gram in the LM3 was divided by the number of different OOVs in the test-set and used as the word n-gram of each OOV word.

4.3. Acoustic modeling

Phone HMMs were adapted to the test-set speaker based on the MAP and the VFS techniques using the training set and AM1 described in Section 3 as the initial model (AM3).

4.4. Experimental results

As described above, the 117 utterances made by a single politician in the discussion noted were used for evaluation. The test-set perplexity and the OOV rate using bigrams and trigrams respectively are given in Table 4. Since the test-set

Table 4: Test-set perplexity and OOV rate for the discussion speech

Language model	Bigram	Trigram	OOV(%)
LM1	1098.0	528.4	9.6
LM2	163.7	151.4	4.8
LM3	143.7	141.4	3.1
LM4	-	-	0
LM5	-	-	0.4
LM6	-	-	0

perplexity for the language models including OOVs using the various methods is not meaningful for comparison with those before adding the OOVs, the test-set perplexity after adding the OOVs is not shown in the table. The table indicates that LM2 and LM3 are more useful than LM1 similarly to the results given in the previous section.

Table 5 presents the results of the recognition experiments. As can be seen, all of the methods investigated in this section are effective in reducing the error rate, and the error rate using LM6 is 28% lower than that using LM1.

Table 5: Results of recognition experiments

Language model	%Corr(%)	Acc(%)	%Err(%)
LM1	44.7	37.6	62.4
LM2	56.9	49.0	51.0
LM3	59.5	49.8	50.2
LM4	59.8	51.8	48.2
LM5	59.6	51.8	48.2
LM6	61.0	52.6	47.4

4.5. Discussion

Similarly to the recognition of presentation speech, it was found that the language model made from the transcribed lectures was more effective than the language model constructed using the broadcast news texts. All of the simple methods for including OOV words in the language model investigated in this paper were verified as being almost equally effective. Among the three methods, the LM6 method using the OOV class trigram divided by the number of different OOV words in the test set utterances was found to be the most effective.

5. CONCLUSION

This paper first introduced a 5-year Japanese national project on spontaneous speech corpus and processing technology started in 1999, and then reported preliminary recognition experiments for recognizing spontaneous speech performed at Tokyo Institute of Technology. The project is being conducted toward realizing three major themes: 1) building a large-scale spontaneous speech corpus consisting of roughly 7M words having a total speech length of 1000 hours, 2) acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information in speech, and 3) constructing a prototype of a spontaneous speech summarization system.

The preliminary recognition experiments have been performed using one speaker's presentation utterances and those made by a politician during a discussion. Recognition results showed that the language models generated from the following methods were effective for recognizing spontaneous speech: a) building a language model by transcribed lectures collected from the WWW, b) adding filled pauses to the corpus based on the statistical characteristics of the filled pauses observed for each speaker, and c) including out-of-vocabulary words (new words) to the language model by several simple methods.

Since the word error rates for these tasks are still very high, it is imperative to collect a large corpus of spontaneous speech and use it for building language and acoustic models to improve the recognition performance. Future research issues include: a) how to transcribe spontaneous speech; b) how to apply morphological analysis to the transcribed spontaneous speech; c) how to build precise and yet general filled pause models (in this paper, they were given only at sentence beginnings and before and after commas in the corpus); d) how to incorporate repairs, hesitations, repetitions, partial words, and disfluency; and e) how to adapt the language models to each task. It is also important to investigate the method of building acoustic models that fit spontaneous speech.

Indispensable in the processing of spontaneous speech will be a paradigm shift from speech recognition to understanding, where underlying messages of the speaker, namely meaning/context that the speaker intends to convey, are extracted instead of transcribing all of the spoken words [3]. Speech summarization, which is one of the main targets of the national project, is considered to be one of the variations of fostering speech understanding [4]. Speech summarization will also be applicable to a range of applications, such as preparing minutes of meetings, close captioning of broadcast news, and presenting information in news-on-demand systems.

ACKNOWLEDGMENT

The authors wish to express their thanks to Professor Tatsuya Kawahara at Kyoto University for his contribution to the national project and for several valuable comments and fruitful discussions related to the preliminary recognition experiments reported in this paper. The authors would also like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news and discussion database.

REFERENCES

[1] K. Maekawa, H. Koiso, S. Furui and H. Isahara: "Spontaneous speech corpus of Japanese," Proc. 2nd International Conference on Language Resources and Evaluation, Athens, Greece, pp. 947-952 (2000).

[2] K. Kato, A. Lee and T. Kawahara: "Topic-independent language model and its adaptation for dictation of lecture speech," IPSJ Technical Report on Spoken Language Information Processing, 26-2 (1999) (in Japanese).

[3] S. Furui: "Steps toward natural human-machine communication in the 21st century," Proc. COST249 Workshop, "Voice Operated Telecom Services," Gent, Belgium, pp. 17-24 (2000).

[4] C. Hori and S. Furui: "Automatic speech summarization based on word significance and linguistic likelihood," Proc. ICASSP2000, Istanbul, Turkey, pp. 1579-1582 (2000).