論文 / 著書情報
Article / Book Information

Title	Towards Automatic Transcription of Spontaneous Presentations
Authors	Takahiro Shinozaki, Chiori Hori, Sadaoki Furui
Citation	Eurospeech 2001, Vol. 1, No. , pp. 491-494
Pub. date	2001, 9

Towards Automatic Transcription of Spontaneous Presentations

Takahiro Shinozaki, Chiori Hori and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
{staka, chiori, furui}@furui.cs.titech.ac.jp

Abstract

This paper reports various investigations on recognizing spontaneous presentation speech in connection with the "Spontaneous Speech" national project started in 1999. Presentation speech uttered by 10 male speakers of approximately 4.5 hours duration has been recognized. Experimental results show that acoustic and language modeling based on an actual spontaneous speech corpus is far more effective than conventional modeling based on read speech. The recognition accuracy has a wide speaker-to-speaker variability according to the speaking rate, the number of fillers, the number of repairs, etc. It was confirmed that unsupervised speaker adaptation of acoustic models was effective to improve the recognition accuracy. The recognition accuracy for spontaneous speech is, however, still rather low, and there remains a large number of research issues.

1. Introduction

Improving the recognition performance for spontaneous speech is crucial to effectively broadening the application of speech recognition. Applying acoustic and language models based on written language to spontaneous speech results in poor recognition accuracy due to the acoustic and linguistic mismatch.

To build models and technology for spontaneous speech recognition, the Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" was started in 1999 under the supervision of Furui [1]. The project will be conducted over a 5-year period in pursuit of the following three major goals:

1) Building a large-scale spontaneous speech corpus consisting of approximately 7M words with a total speech length of 800 hours. The majority of the recordings will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. Since there is no clear definition of words in Japanese and no spacing between words in written Japanese sentences, a morphological analysis program will be used to split transcribed sentences into morphemes.

2) Acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information.

3) Constructing a prototype of a spontaneous speech summarization system.

This paper reports results of preliminary experiments utilizing a part of the corpus that has so far been built. Section 2 describes the task and experimental conditions. Section 3 describes acoustic and language models for recognition. Recognition results are presented in Section 4, and Section 5 gives some analysis on individual variations of recognition results. Section 6 reports the improvement by unsupervised speaker adaptation. Finally some conclusions are given in Section 7.

2. Recognition task and experimental conditions

2.1. Recognition task

Presentation speech uttered by 10 male speakers was used as a test set of speech recognition. Table 1 shows an outline of the test set. The top four presentations in the table were on the subject of speech.

Morphemes (which will be called "words" hereafter in this paper) were used as units for statistical language modeling. For all the following recognition performances, word-based performance is measured. Fillers are counted as words and taken into account in calculating the accuracy.

Table 1: Recognition test set of presentations

ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
P25	Phonetics Soc. Jap.	27
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Corpus Meeting	14
Y05	Spont. Speech Corpus Meeting	15

2.2. Experimental conditions

Sounds were digitized with 16kHz sampling and 16bit quantization. They were segmented into utterances using silence periods longer than 500ms. Feature vectors had 25 elements consisting of 12 MFCC, their delta, and delta log energy. CMS (cepstral mean subtraction) was applied to

each utterance. HTK v2.2 [2] was utilized for acoustic modeling and speaker adaptation. Language models were made by the use of CMU SLM Tool Kit v2.05. The Julius v3.1 decoder [3] was used for speech recognition. Language model weighting used in the decoding process was changed for each combination of acoustic and language model but kept constant over all speech in the test set.

3. Language and acoustic modeling

3.1. Corpora

The following two corpora were used for training.

• Spontaneous Speech Corpus (SSC): A part of the corpus completed by the end of December 2000, consisting of approximately 1.5M words of transcriptions, was used. The training set consisted of 610 presentations; 274 academic conference presentations and 336 simulated presentations. The simulated presentations talking about a wide variety of topics including the subjects' experiences in their daily lives were specially recorded for the project.

• Web corpus: Transcribed presentations having roughly 76k sentences with 2M words were collected from the World Wide Web. Spontaneous speech usually includes various filled pauses but they are not included in this presentation corpus. An effort was thus made to add filled pauses to the presentation corpus based on the statistical characteristics of the filled pauses. Their topics covered wide domains including social issues and memoirs.

3.2. Language modeling

The following three language models were built. Each model consisted of bigrams and reverse trigrams with backing-off. Their vocabulary sizes were all 30k.

SpnL: Made using the 610 presentations in the SSC. The speakers had no overlap with those of the test set. Since there were no punctuation marks in the transcription, commas were inserted at silences of 200ms or longer duration.

WebL: Made using the text of our Web corpus.

WebSpL: Made by adding whole text of a textbook on speech processing authored by Furui to the Web corpus with equal weighting for task adaptation. The textbook contains about 63k words.

Table 2 shows an outline of the language models.

Table 2: Corpus size for training each language model

Language model	Corpus size [words]
SpnL	1.5 M
WebL	2 M
WSpL	2+0.06 M

3.3. Acoustic modeling

The following two tied-state triphone HMMs were made. Both models have 2k states and 16 Gaussian mixtures in each state.

SpnA: Using 338 presentations in the SSC uttered by male speakers (approximately 59 hours). The speakers had no overlap with those in the test set.

RdA: The acoustic model made by Information-technology Promotion Agency (IPA) and contained in the CD-ROM "Japanese Dictation Toolkit 99". Approximately 40-hour long read speech uttered by many speakers was used.

4. Experimental results

4.1. Test-set perplexity and OOV rate

Figure 1 presents test-set perplexity of tri-grams and out-of-vocabulary (OOV) rate for each language model. Perplexity of **SpnL** made from the SSC, is clearly better than that of other models. **WebL** indicates high perplexity and OOV rate. This is because **WebL** is edited as a text and their topics are general. OOV rate of **WSpL** is smaller than that of **WebL** for the four left-hand-side speeches. This shows that task adaptation by adding the textbook worked well. **SpnL** is superior to **WSpL** also in terms of OOV rate.

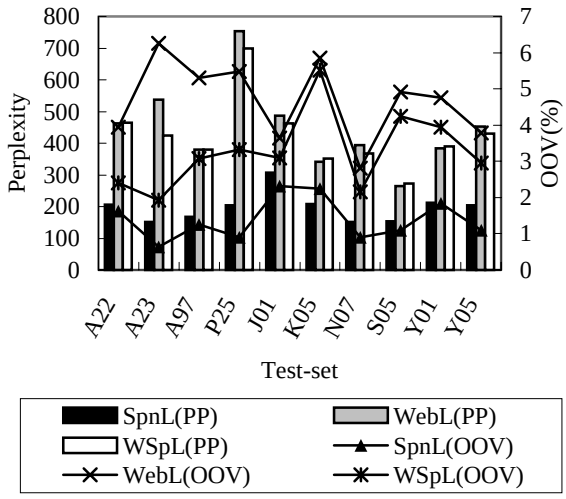

Figure 1. Test-set perplexity and OOV rate for the three language models.

4.2. Effects of language modeling

Figure 2 shows recognition results for the three language models when **SpnA** is used as the acoustic model. **SpnL** achieves the best results. **WSpL** achieves better results than **WebL**, especially for test sets A22, A23, A97 and P25, reflecting the test-set perplexity and OOV rate reduction. Mean accuracies are 64.3%, 54.9% and 57.1% for **SpnL**, **WebL** and **WSpL**, respectively. A supplementary

experiment was performed to analyze the effects of OOV rate and test set perplexity to the accuracy. In this experiment, OOV words were added to the language models as "unknown" class words; 489 words and 710 words were added to **SpnL** and **WSpL**, respectively. Resulting mean word accuracies using **SpnL** and **WSpL** were 65.8% and 59.9%, respectively. These results indicate that OOV is an equally import problem as test-set perplexity in these models.

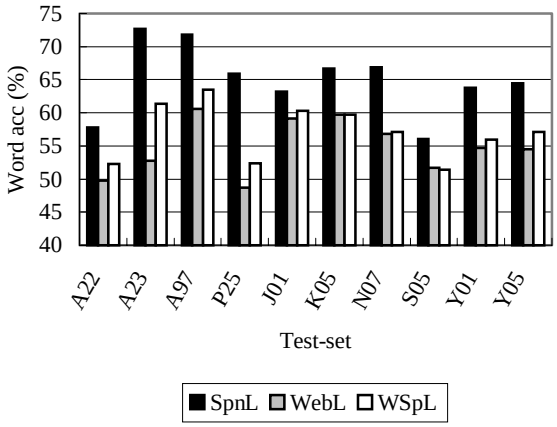

Figure 2. Word accuracy for the three language models.

4.3. Effects of acoustic modeling

The recognition results for **SpnA** and **RdA** when **SpnL** is used as the language model are shown in Fig. 3. Mean accuracies are 64.3% and 53.0% for **SpnA** and **RdA**, respectively. **SpnA** made from the SSC achieves much better results than **RdA** made from read speech. This is probably because **SpnA** has better coverage of triphones and better matching of acoustic characteristics corresponding to the speaking style. **SpnA** also has better matching of recording conditions with the test set. SSC and IPA corpora were both recorded used close-talking microphones, but types of the microphones were different.

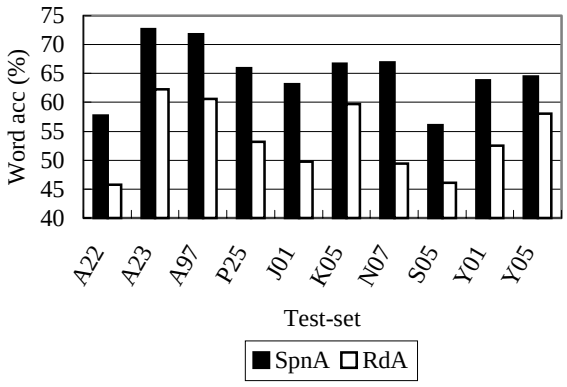

Figure 3. Word accuracy for the two acoustic models.

5. Individual differences

As shown in Figs. 2 and 3, the word accuracy largely varies from speaker to speaker. There exist many factors that affect the accuracy of spontaneous speech recognition. These factors include individual voice characteristics and speaking manner including noises like cough. Although all utterances were recorded using the same close-talking microphones, acoustic conditions still varied according to the recording environment.

Figure 4 presents relationship between speaking rate and word accuracy when **SpnL** and **SpnA** were used as language and acoustic models. The speaking rate was calculated using actual speech periods after removing pauses. 10 dots in the figure correspond to individual speakers. A MMSE line fitted to those dots is also shown in the figure. The correlation coefficient is –0.58. Faster speech generally produces more errors.

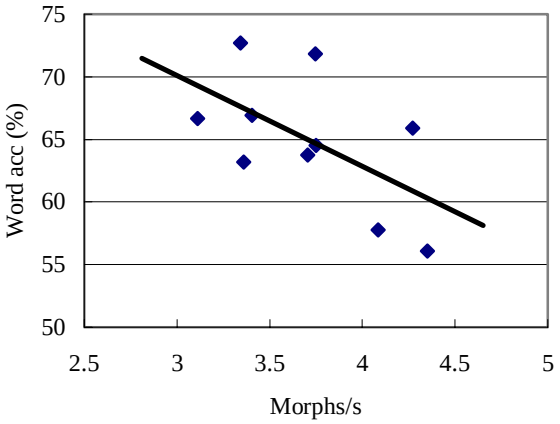

Figure 4. Speaking rate vs. word accuracy.

Figures 5 and 6 respectively show the effects of frequencies of fillers and repairs on word accuracy. The recognition conditions were the same as those for Fig. 4. There is a general tendency that the more frequently the filler and/or the repair occurs, the more recognition error occurs.

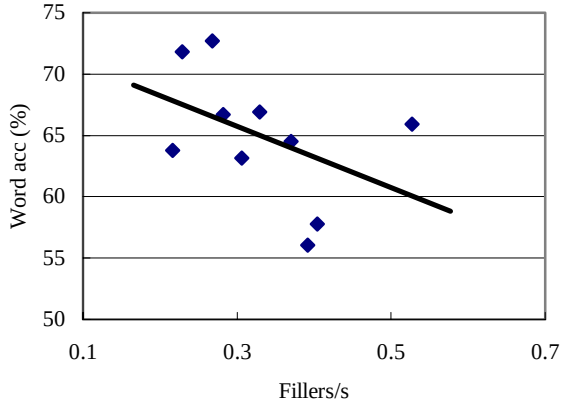

Figure 5. Filler frequency vs. word accuracy.

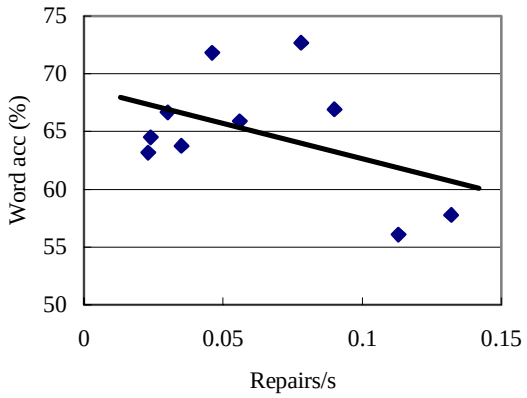

Figure 6. Repair frequency vs. word accuracy.

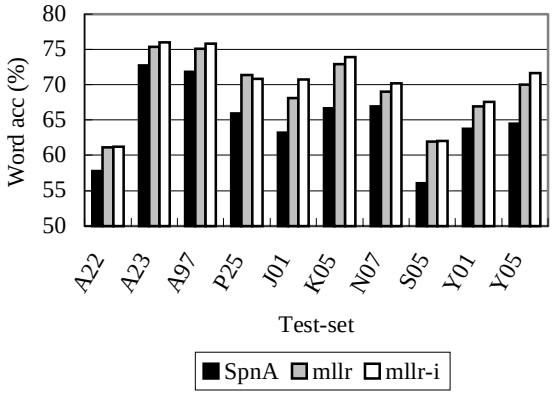

Figure 7. Results of unsupervised adaptation.

6. Unsupervised adaptation

A batch-type unsupervised adaptation method was incorporated to cope with the speech variation due to speakers and recording environment. We applied the MLLR method using a binary regression class tree to transform Gaussian mean vectors. The regression class tree was made using a centroid-splitting algorithm. The actual classes used for transformation were determined on run time according to the amount of data assigned to each class.

The following steps were carried out. The adaptation was performed based on recognition results and no confidence measure was applied.

1. Making a regression class tree having 64 leaf nodes for the **SpnA** phone model.

2. Recognizing the test-set utterances using the **SpnA** as a speaker independent model.

3. Applying the MLLR adaptation based on the recognition result for each utterance to make a speaker adaptive model.

4. Re-recognizing the test-set utterances using the speaker adaptive model.

5. Iterating the adaptation process using the resulting transcription.

Figure 7 presents the effect of the adaptation when **SpnL** was used as the language model. "SpnA" indicates the baseline condition. "mllr" indicates the result without iterations and "mllr-i" indicates the results after one iteration of adaptation. The single step of MLLR improved word accuracy by 2 to 6 %, and the second adaptation step further improved accuracy by 1% in average. By applying the two steps of MLLR, error rate was reduced by 15% relative to the speaker independent case.

7. Conclusions

This paper reported experimental results for recognizing spontaneous presentation speech. Language models based on a spontaneous speech corpus and Web corpus were compared in terms of test-set perplexity, OOV rate, and word (morpheme) accuracy. Two acoustic models made by spontaneous speech and read speech were also compared. Both comparisons showed that models made from spontaneous speech were superior to models based on read speech. It was revealed that the recognition accuracy had a wide speaker-to-speaker variability. Correlation between word accuracy and speaking rate, filler and repair frequency was observed. When linguistic and acoustic models made form spontaneous speech were used, an average word recognition accuracy of 64.3% was achieved. This performance improved to 69.8% with the help of unsupervised MLLR adaptation for the acoustic model.

Since word accuracy for this task is still very low, further improvement is indispensable for building application systems. Future research issues include a) how to transcribe and annotate spontaneous speech, b) how to build filled pause models, and c) how to incorporate repairs, hesitations, repetitions, partial words, and disfluencies. Adaptation for speaking styles and topics of presentations is also crucial.

8. Acknowledgment

The authors wish to express their thanks to the members of "Spontaneous Speech" national project for constructing the corpus and for fruitful discussions.

9. References

[1] S. Furui, et al., "Toward the realization of spontaneous speech recognition – Introduction of a Japanese Priority Program and preliminary results", Proc. ICSLP, Beijing, pp. 518-521, 2000

[2] S. Young, et al., "The HTK Book, Version 2.2", Entropic Ltd, 1999.

[3] A. Lee, et al., "An Efficient Two-pass Search Algorithm using Word Trellis Index", Proc. ICSLP, Australia, pp.1831-1834, 1998