

論文 / 著書情報
Article / Book Information

Title	Recent advances in Japanese broadcast news transcription
Authors	Katsutoshi Ohtsuki, Sadaoki Furui, Naoyuki Sakurai, Atsushi Iwasaki, Zhi-Peng Zhang
Citation	Eurospeech '99, Vol. 2, No. , pp. 671-674
Pub. date	1999, 9

RECENT ADVANCES IN JAPANESE BROADCAST NEWS TRANSCRIPTION

Katsutoshi Ohtsuki¹, Sadaoki Furu², Naoyuki Sakurai², Atsushi Iwasaki², Zhi-Peng Zhang²

¹NTT Cyber Space Laboratories, Media Processing Project

1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan

²Tokyo Institute of Technology, Department of Computer Science

2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan

ABSTRACT

In this paper, we report on language modeling and acoustic modeling studies for Japanese broadcast news speech recognition. We constructed a language model that reduces recognition errors by utilizing context-dependent readings of Japanese characters. We also introduced filled-pause modeling into the language model. To improve the model's performance for a series of sentences spoken by one speaker, an on-line incremental speaker adaptation was combined with automatic detection of speaker changes. By incorporating all the above methods, we achieved a 25.1% reduction in word error rate over the baseline results. This paper also reports on our preliminary studies on topic extraction and summarization of broadcast-news speech.

1. INTRODUCTION

We have been developing a large-vocabulary continuous-speech recognition (LVCSR) system for Japanese broadcast-news speech transcription. This is a part of a joint research project with NHK broadcasting whose goal is the closed-captioning of TV programs. In this paper, we report on recent progress in both language modeling and acoustic processing.

Japanese text includes a mixture of three kinds of characters: Chinese characters (Kanji) and two kinds of Japanese characters (Hira-gana and Kata-kana). Each Kanji has multiple readings, and correct readings can only be decided according to context. Conventional language models are built using the written forms of words, and usually equal probability is assigned to all possible readings of each word. This method causes recognition errors in Japanese speech recognition because the assigned probability is sometimes very different from the true probability. We therefore constructed a language model that depends on the readings of words in order to take into account the frequency and context-dependency of the readings.

Broadcast news speech includes filled-pauses at the beginning and in the middle of sentences, which cause recognition errors in our previous language models that use news manuscripts written prior to broadcasting. To cope with this problem, we introduced filled-pause modeling into the language model.

Since, in broadcast news, each speaker utters several sentences in succession, the recognition error rate can be reduced by incrementally adapting the acoustic models within a segment that contains only one speaker. We

applied on-line, unsupervised, instantaneous and incremental speaker adaptation combined with automatic detection of speaker changes.

2. JAPANESE LVCSR SYSTEM

2.1 Baseline Language Models

The broadcast-news manuscripts that were used for constructing the language models were taken from NHK news broadcasts over a period between July 1992 and May 1996, and comprised roughly 500k sentences and 22M words. To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words. A word-frequency list was derived for the news manuscripts, and the 20k most frequently used words were selected as vocabulary words. This 20k vocabulary covered approximately 98% of the words in the broadcast-news manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using Katz's back-off smoothing method.

2.2 Acoustic Models

The feature vector consisted of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector was 34. Cepstral coefficients were normalized by the cepstral mean subtraction (CMS) method.

The acoustic models were gender-dependent shared-state triphone HMMs and were designed using tree-based clustering. They were trained using phonetically-balanced sentences and dialogues read by 53 male speakers and 56 female speakers. The contents were completely different from the broadcast-news task. The total number of training utterances was 13,270 for male and 13,367 for female, and the total length of the training data was approximately 20 hours for each gender. The total number of HMM states was approximately 2,000 for each gender, and the number of Gaussian mixture components per state was four.

2.3 Evaluation Data

News speech data, from TV broadcasts in July 1996, were divided into two parts, a clean part and a noisy part, and were separately evaluated. The clean part consisted of utterances with no background noise, and the noisy part consisted of utterances with background noise. The noisy part included

spontaneous speech such as reports by correspondents. We extracted 50 male utterances and 50 female utterances for each part, yielding four evaluation sets: male-clean (m/c), male-noisy (m/n), female-clean (f/c), female-noisy (f/n). Each set included utterances by five or six speakers. All utterances were manually segmented into sentences.

3. IMPROVEMENTS IN LANGUAGE MODELS

3.1 Reading Variety of Japanese Characters

In Japanese sentences, Chinese characters (Kanji) can be read in various ways and the way of reading a Kanji is decided according to context. Conventional language models depend on the written form of a word and assign equal probability to all possible readings of a word. These models cause recognition errors in Japanese speech recognition because the assigned probability is sometimes very different from the true probability. Last year we proposed a language model in which reading probabilities were multiplied with n-gram probabilities [1]. The improvement in recognition performance with this language model was limited since contexts of words were not taken into consideration.

3.2 Reading-Dependent Language Model

We have constructed a new language model in which a word with multiple readings is split into different language model entries according to its readings. Table 1 shows the OOV rates of the 20k vocabulary of the conventional language model (LM1) and that of the reading-dependent language model (LM2) for the training text data and the evaluation sets. The reading-dependent language model only slightly increases the OOV rate.

Table 1: OOV rates for 20k vocabulary size [%]

Language model	Training text	Evaluation sets			
		m/c	m/n	f/c	f/n
LM1	2.27	0.81	2.88	1.21	3.49
LM2	2.39	0.86	3.02	1.26	3.68

3.3. Filled-Pause Modeling

Broadcast news speech includes filled-pauses at the beginning and in the middle of sentences. These pauses caused recognition errors in our previous language models, which were built using the original news manuscripts. After analyzing the transcribed text of broadcast news, we found that major filled-pauses were /e/ and /eh/, and they often appeared at the beginning of sentences or just after commas (the probabilities of their occurrences are shown in Table 2). Whereas our baseline language model was constructed after removing all punctuation marks from the original training text, the new language model had training text that included commas and had three acoustic models for each comma and sentence beginning: silence, /e/, and /eh/.

Table 2: Probabilities of filled-pauses in the transcribed broadcast news [%]

	Sentence beginning	After comma
/e/	12.1	3.4
/eh/	2.1	1.0

3.4 Experimental Results

Table 3 shows the experimental results for the baseline language model (LM1) and the new language models. LM2 was the reading-dependent language model, and LM3-1 was a modification of LM2 that kept commas in the training text but gave only silence as the acoustic model to sentence beginnings and commas. LM3-2 was a modification of LM3-1 in which the three acoustic models were given to sentence beginnings and commas. LM2 (trigram) reduced the word error rate by 4.4 % relative to LM1, averaged over the all evaluation sets, despite its slightly higher OOV rate. By considering commas (LM3-1), the recognition performances were improved, and there was further improvement after adding pronunciations, /e/ and /eh/, to the sentence beginnings and commas (LM3-2). LM3-2 (trigram) model reduced the word error rate by 7.9 % relative to the results of LM2, averaged over all the evaluation sets,.

Table 3: Experimental results with various language models (word error rate [%])

Language model		Evaluation sets			
		m/c	m/n	f/c	f/n
bigram	LM1	20.9	40.3	18.3	45.2
	LM2	20.4	39.1	17.3	42.7
	LM3-1	19.9	37.8	16.9	42.3
	LM3-2	18.0	37.2	16.1	41.7
trigram	LM1	17.6	37.2	14.3	41.2
	LM2	16.8	35.9	13.6	39.3
	LM3-1	16.7	34.6	13.7	39.1
	LM3-2	14.2	33.1	12.9	38.1

4. ON-LINE SPEAKER ADAPTATION

4.1 Speaker Adaptation based on Speaker Change Detection

In broadcast news, each speaker usually utters several sentences in succession. Therefore, incremental acoustic model adaptation within a segment in which one speaker utters several sentences is expected to be effective. We applied on-line unsupervised speaker adaptation in conjunction with automatic speaker-change detection to successive incoming broadcast news utterances. The maximum likelihood linear regression (MLLR) [2] and

vector-field smoothing (VFS) [3] methods were instantaneously and incrementally carried out for each utterance.

The adaptation process is as follows. For the first input utterance, the speaker-independent model is used for both recognition and adaptation, and the first speaker-adapted model is created. For the second input utterance, the likelihood value of the utterance given the speaker-independent model and that given the speaker-adapted model are calculated and compared. If the former value is larger, the utterance is considered to be from a new speaker, and another speaker-adapted model is created. Otherwise, the existing speaker-adapted model is incrementally adapted. For the succeeding input utterances, speaker changes are detected in the same way by comparing the acoustic likelihood values obtained from the speaker-independent model and the speaker-adapted models. If the speaker-independent model yields a larger likelihood than any of the speaker-adapted models, a speaker change is detected and a new speaker-adapted model is constructed. In the experiment, to take advantage of our broadcast-news structure and to reduce the computational time, the two most recently constructed speaker-adapted models were kept and older models were removed. Figure 1 is a flow chart of the speaker adaptation process.

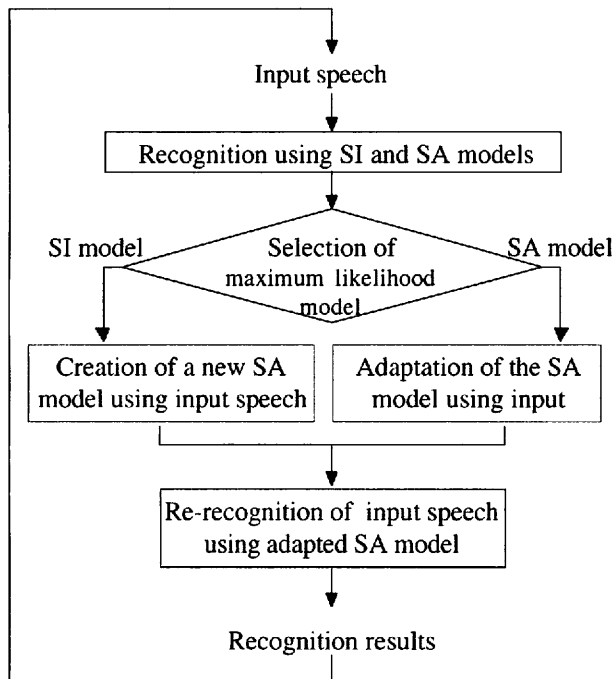


Figure 1: On-line speaker adaptation process (SI: speaker independent, SA: speaker adapted)

4.2 Experimental Results

We tried to determine the appropriate number of transformation matrices, or clusters, for MLLR. MLLR with a single cluster, two clusters (silence, and all phonemes), three clusters (silence, vowels, and consonants), seven

clusters (silence, consonants, and five Japanese vowels), and nine clusters (silence, five Japanese vowels, and three Japanese consonant categories) were studied using male-clean speech. Table 4 shows recognition performances. “Baseline” were for the LM3-2 language model and no acoustic model adaptation. MLLR with seven clusters achieved the best performance in this experiment.

Table 4: Speaker adaptation performances with several types of MLLR (word error rate [%])

Baseline		18.0
MLLR type	1 cluster	15.6
	2 clusters	16.0
	3 clusters	16.4
	7 clusters	14.7
	9 clusters	15.8

Table 5 shows the results of the speaker adaptation experiments for MLLR with a single cluster and with seven clusters. The performance of the model by MLLR with a single cluster was better than that of speaker-independent models, and further improvement was achieved by using the seven phoneme clusters. The latter method reduced the word error rate by 11.8%, averaged over male and female, relative to the results for the speaker-independent models.

Table 5: Experimental results of speaker adaptation (word error rate [%])

Language model	MLLR type	Evaluation sets		
		m/c	f/c	average
bigram (LM3-2)	Baseline	18.0	16.1	17.0
	1 cluster	15.6	15.8	15.6
	7 clusters	14.7	14.4	14.5
trigram (LM3-2)	Baseline	14.2	12.9	13.5
	1 cluster	12.7	12.5	12.6
	7 clusters	12.1	11.8	11.9

4.3 Detected Speaker Changes vs. Given Speaker Changes

For comparison with the above experiments, we conducted a speaker adaptation experiment in which correct speaker changes were given. In this experiment, the speaker-independent model was used as the initial model for adaptation, irrespective of the likelihood value, every time the speaker changed. The experimental results are shown in Table 6. In the speaker adaptation experiments, lower word error rates were obtained with automatically detected speaker changes than with given speaker changes. This result shows that the acoustic model which has the highest likelihood value with respect to the input speech should be used as the initial model for speaker adaptation.

Table 6: Comparison of speaker adaptation results with detected and given speaker changes (word error rate [%])

Language model	MLLR type	Speaker changes	Evaluation sets		
			m/c	f/c	average
bigram (LM3-2)	1 cluster	given	16.0	15.5	15.8
		detected	15.6	15.8	15.7
	7 clusters	given	15.7	14.3	15.0
		detected	14.7	14.4	14.5
trigram (LM3-2)	1 cluster	given	12.7	13.0	12.9
		detected	12.7	12.5	12.6
	7 clusters	given	12.3	11.9	12.1
		detected	12.1	11.8	11.9

5. Topic-word Extraction and Summarization

Summarizing transcribed news speech is useful for retrieving or indexing broadcast news. Since the speech recognition process is error prone, especially for function words which are usually short and obscure in speech, the conventional summarization techniques for written sentences cannot be applied to the automatically transcribed speech. Therefore, we have been investigating the following summarization method that relies on topic-word extraction,

1. *Extract topic-words from speech recognition results.*
2. *Connect adjacent topic-words into compound words,*
3. *Insert a verb between two topic-words, and*
4. *Insert postpositional particles after topic-words and verbs.*

We extracted the topic-words from nouns in the speech recognition results by using a method based on a significance measure [1][4]. The extracted topic-words were compared with 'true' topic-words, which were given by three human subjects. Experimental results showed that, when the top five topic-words were chosen (recall = 13%), 87% of them were correct. Based on these topic-words, summarizing sentences were created by the above process. This process is still being evaluated, but its preliminary results are promising.

6. SUMMARY

This paper reported on recent advances in Japanese broadcast news transcription. While some of the problems that we investigated were Japanese-specific, others are language-independent. One of the Japanese-specific problems we investigated is the variety of ways that exist to read Chinese (Kanji) characters. Conventional language models depend on written forms of words and assign equal probability to all possible readings of a word. These models cause recognition

errors because the assigned probability is sometimes very different from the true probability. We constructed a language model that took into account the frequency and context-dependency of the readings of Kanji characters. This method reduced the word error rate by 4.4%.

To cope with the recognition errors caused by filled-pauses, we introduced filled-pause modeling into the language model. This reduced the word error rate by 7.9%.

We applied on-line, unsupervised, instantaneous and incremental speaker adaptation to successive utterances spoken by the same speaker combined with automatic detection of speaker changes. The maximum likelihood linear regression (MLLR) method with seven clusters combined with vector-field smoothing (VFS) technique reduced word error rate by 11.8%.

By incorporating all the above methods, we achieved an 11.9% word error rate averaged over males and females for clean parts of broadcast news speech, which was a 25.1% reduction in word error rate over the baseline results.

This paper also reported on our preliminary experiments on broadcast-news summarization.

ACKNOWLEDGMENTS

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database. The authors are also grateful to Nihon Keizai Shimbun Incorporated for allowing us to use the newspaper text database (Nikkei CD-ROM 90-94) for our research. The authors also would like to thank Dr. T. Fuchi at NTT Cyber Space Laboratories for letting us use their morphological analysis program. This work is supported in part by the International Communications Foundation.

REFERENCES

- [1] Furui, et al., "Japanese Broadcast News Transcription and Topic Detection," Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp. 144-149, 1998.
- [2] C. J. Leggetter et al., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, pp. 171-185, 1995-9.
- [3] K. Ohkura et al., "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. ICSLP'92, pp. 369-372, 1992.
- [4] K. Ohtsuki et al., "Topic Extraction with Multiple Topic-words in Broadcast-news Speech," Proc. ICASSP'98, pp. I-329-332, 1998.