論文 / 著書情報
Article / Book Information

Title	Improvements in Japanese broadcast news transcription
Authors	Katsutoshi Ohtsuki, Sadaoki Furui, Naoyuki Sakurai, Atsushi Iwasaki, Zhi-Peng Zhang
Citation	DARPA Broadcast News Workshop, Washington, Vol. , No. , pp. 231-236
Pub. date	1999, 2

Proc.
DARPA Broadcast News Workshop, Washington, D.C. (1999) pp. 231-236

Improvements in Japanese Broadcast News Transcription

Katsutoshi Ohtsuki[†], *Sadaoki Furui, Naoyuki Sakurai, Atsushi Iwasaki and Zhi-Peng Zhang*

Tokyo Institute of Technology, Department of Computer Science
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan / furui@cs.titech.ac.jp
[†]NTT Cyber Space Laboratories
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan / ohtsuki@nttspch.hil.ntt.co.jp

ABSTRACT

This paper reports on recent improvements in Japanese broadcast news transcription and topic extraction. We constructed a language model that depends on the readings of words in order to prevent recognition errors caused by context-dependent readings of Japanese characters. We also introduced interjection 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 applied. We investigated a method for extracting topic-words from the speech recognition results that was based on a significance measure. This paper also proposes a new formulation for speech recognition/understanding systems, in which the a posteriori probability of a message that the speaker intends to address given an observed acoustic sequence is maximized. We applied the formulation to rescoring the recognition hypotheses.

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 with NHK broadcast company whose goal is the creation of LVCSR technology for automatically generating closed-captions of TV programs. In this paper, we report on recent progress in both language modeling and acoustic processing.

Japanese text is written by 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 written forms of words, and usually equal probability is assigned to all possible readings of each word. This causes recognition errors 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 interjections at the beginning and in the middle of sentences, which cause recognition errors in our language models that use news manuscripts written prior to broadcasting. To cope with this problem, we introduced interjection modeling into the language model.

Since, in broadcast news, each speaker utters several sentences in succession, the recognition error rate can be reduced by adapting acoustic models incrementally 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.

Summarizing transcribed news speech is useful for retrieving or indexing broadcast news. We investigated a method for extracting topic words from nouns in the speech recognition results on the basis of a significance measure. The extracted topic-words were compared with "true" topic-words, which were given by three human subjects.

This paper also proposes a new formulation for speech recognition/understanding systems, in which the a posteriori probability of a message that the speaker intends to address given an observed acoustic sequence is maximized as an extension of the current criterion that maximizes the probability of a word sequence hypothesis. We applied the formulation to rescoring the recognition hypotheses.

2. JAPANESE LVCSR SYSTEM

2.1 Baseline Language Models

The broadcast-news manuscripts that were used for constructing the language models were taken from the 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 covers about 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 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 4.

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, male-noisy, female-clean, female-noisy. 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 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 training text data and 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. Interjection Modeling

Broadcast news speech includes interjections at the beginning and in the middle of sentences. They cause recognition errors since our language models, which were built using news manuscript written prior to broadcasting, can not handle those interjections. Analyzing the transcribed text of broadcast news, we found that major interjections were /e/ and /eh/, and they often appeared in the beginning of sentences and just after commas with probabilities 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 interjections 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 is the reading-dependent language model, and LM3-1 is a modification of LM2 by keeping commas in the training text but giving only silence as the acoustic model to sentence beginnings and commas. LM3-2 is a modification of LM3-1 in which the three acoustic models are given to sentence beginnings and commas. LM2 reduced the word error rate by 4.7 % relative to LM1 despite its slightly higher OOV rate. By considering commas (LM3-1), the performances were improved, and there was further improvement after adding pronunciations, /e/ and /eh/, to the sentence beginnings and commas (LM3-2). LM3-2 model reduced the word error rate by 10.9 % relative to the results of LM2.

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

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 MLLR [2] -MAP [3] and VFS (vector-field smoothing) [4] 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 the beginning of 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 of each utterance obtained from the speaker-independent model and some 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 our experiment, to take advantage of our broadcast-news structure and to reduce the computational time the two most recently constructed speaker-adapted models are kept and older models are removed.

Two types of the MLLR method were examined. First, all phoneme models were adapted using a common linear regression expression. Next, phonemes were divided into 7 clusters (silence, consonants, and five Japanese vowels) and converted using individual expressions. Table 4 shows the results of the speaker adaptation experiments. "Baseline" indicates the results for the LM3-2 language model and no acoustic model adaptation.

The adaptation using MLLR with a single cluster improved the performance, and further improvement was achieved by using the seven phoneme clusters. The latter method reduced the word error rate by 11.8 % relative to the results for the speaker-independent models.

Table 4: Experimental results of speaker adaptation (word error rate [%])

Language model	MLLR type	Evaluation sets		
		m/c	f/c	average
bigram	Baseline	18.0	16.1	17.0
	1 cluster	15.5	15.8	15.6
	7 clusters	14.7	14.4	14.5
trigram	Baseline	14.2	12.9	13.5
	1 cluster	12.7	12.5	12.6
	7 clusters	12.1	11.8	11.9

5. TOPIC EXTRACTION

Transcribed broadcast news speech can also be used for automatic indexing or summarizing of the news. We have been investigating methods for extracting a set of topic words expressing the content of each broadcast news automatically from each news [1][5]. Topic-words are extracted on the basis of a significance score calculated as follows [6].

$$SgScore(w_i) = g_i \cdot \log \frac{G_A}{G_i} \quad (i = 1,2,...,N) \qquad (1)$$

where g_i is the number of occurrences of a word w_i in a news article, G_i is the frequency of the word w_i in all the training news articles, and G_A is the summation of all G_i's. To calculate the G_i's and G_A values, we used news articles from roughly five years of the Nikkei business newspaper.

We applied the topic-extraction method to the transcribed speech that was obtained using the methods described in the previous sections. The extracted topic-words were compared with "true" topic-words which were given by three human subjects. The results are shown in Figure 1. When the top five topic-words were chosen (recall=13%), 87% of them were correct on average.

6. MESSAGE-DRIVEN SPEECH RECOGNITION

6.1 A Communication Model

State-of-the-art automatic speech recognition systems employ the criterion of maximizing $P(W|X)$, where W is a word sequence, and X is an acoustic observation sequence. This criterion is reasonable for dictating read speech. However, the ultimate goal of automatic speech recognition is to extract the underlying messages of the speaker from the

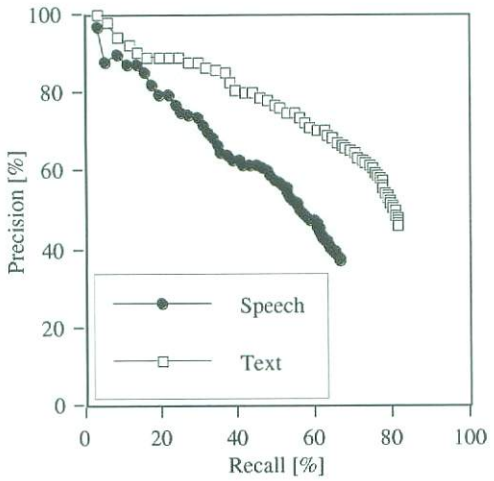

Figure 1: Topic extraction results

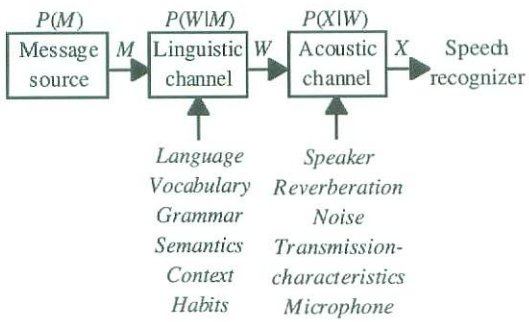

Figure 2: A communication - theoretic view of speech generation and recognition

speech signals. Hence we need to model the process of speech generation and recognition as shown in Fig. 2 [7], where M is the message (content) that a speaker intended to convey.

According to this model, the speech recognition process is the problem of estimating M to maximize $P(M|X)$. We propose a new formulation to solve this problem [8]. It covers the various approaches that have been attempted and also suggests new approaches. We consider that the message M is represented by a co-occurrence of words, and based on this consideration, we propose a new formulation for speech recognition. We apply this formulation to broadcast-news speech transcriptions and show how it reduces word error rate.

6.2 Message-Driven Speech Recognition

According to Fig. 2, the speech recognition process is represented as the maximization of the following a posteriori probability,

$$\max_M P(M|X) = \max_M \sum_W P(M|W)P(W|X). \quad (2)$$

Using Bayes' rule, Eq. (2) can be expressed as

$$\max_M P(M|X) = \max_M \sum_W \frac{P(X|W)P(W|M)P(M)}{P(X)}. \quad (3)$$

For simplicity, we can approximate the equation as

$$\max_M P(M|X) \approx \max_{M,W} \frac{P(X|W)P(W|M)P(M)}{P(X)}. \quad (4)$$

$P(X|W)$ is calculated using hidden Markov models in the same way as in usual recognition processes. We assume that $P(M)$ has a uniform probability for all M. Therefore, we only need to consider further the term $P(W|M)$. We assume that $P(W|M)$ can be expressed as follows.

$$P(W|M) \approx P(W)^{(1-\lambda)} P'(W|M)^{\lambda}, \quad (5)$$

where λ, $0 \le \lambda \le 1$, is a weighting factor. $P(W)$, the first term of the right hand side, represents a part of $P(W|M)$ that is independent of M and can be given by a general statistical language model. $P'(W|M)$, the second term of the right hand side, represents the part of $P(W|M)$ that depends on M. The latter term can be represented in various ways, whether the dependency of M is represented explicitly or implicitly. The explicit formulation usually needs to represent M by a finite number of topic classes. Approaches that change language models according to the estimated topic class M (e. g. [9]) correspond to this formulation. Approaches using probabilistic state transition networks [10] or HMM [11] for forming semantic language models are also classified into this category. A cache model [12] is one of the approaches in which M is implicitly represented.

In this paper, we consider that M is represented by a co-occurrence of words based on the distributional hypothesis by Harris [13]. Similar methods include a method that uses a thesaurus for measuring semantic similarity between words and one that clusters words based on some similarity measures. Since these approaches formulate $P'(W|M)$ without explicitly representing M, they can use information about the speaker's message M without being affected by the quantization problem of topic classes.

6.3 Word Co-Occurrence Score

As mentioned above, $P'(W|M)$, the second term on the right hand side of Eq. (5), is represented by word co-occurrences. Since most of the words that express messages or topics are nouns, we extracted only nouns from the N-best hypotheses of the word sequences which were obtained using the trigram language model. Message-driven speech recognition results were obtained by rescoring the hypotheses by adding word co-occurrence scores for every pair of

nouns. The co-occurrence score was calculated based on the mutual information as follows;

$$CoScore(w_i, w_j) = \log \frac{p(w_i, w_j)}{\left(p(w_i)p(w_j)\right)^{1/2}}, \qquad (6)$$

where $p(w_i, w_j)$ is the probability of observing words w_i and w_j in the same news article, and $p(w_i)$ and $p(w_j)$ are the probabilities of observing word w_i and w_j in all the articles, respectively. To compensate the probabilities of the words with very low frequency, a square root term was employed in the denominator of the equation. The co-occurrence scores were calculated using the same database as that was used for language modeling.

6.4 Experimental Results

Table 5 shows the word error rates obtained by rescoring with word co-occurrence scores. The results before rescoring are also shown in Table 5 for comparison. The weighting factor for co-occurrence score, λ, was appropriately set on the basis of preliminary experiments. As shown in Table 5, word error rates for the clean set were reduced by incorporating $P(W|M)$.

Table 5: Comparison of word error rates [%] without or with $P(W|M)$

Language model	Evaluation sets	
	m/c	m/n
$P(W)$	16.8	35.9
$P(W\|M)$	16.1	35.7

7. 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 depended on the readings of words and considered the frequency and context-dependency of the readings. This method reduced the word error rate by 4.7%.

To cope with the recognition errors caused by interjections, we introduced interjection modeling into the language model. This reduced the word error rate by 10.9%.

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 MLLR method with 7 clusters combined with MAP and VFS techniques 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.

We investigated a method for extracting topic-words from nouns in the speech recognition results on the basis of a significance measure. When the top five topic-words were chosen (recall=13%), 87% of them agreed with the topic-words extracted by human subjects.

This paper also proposed a new formulation for speech recognition/understanding systems, in which the a posteriori probability of a message that the speaker intends to address given an observed acoustic sequence is maximized as an extension of the current criterion that maximizes the probability of a word sequence hypothesis. We assumed that a speaker's message is represented by a co-occurrence of words in the utterance, and we employed a co-occurrence score of words measured by mutual information to rescore word sequence hypotheses. Experimental results show that the word error rate of broadcast news transcription is reduced by the method.

ACKNOWLEDGMENTS

The authors wish to express their appreciation of Dr. B.-H. Juang at Bell Labs for several fruitful discussions. 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] S. 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] J. -L. Gauvain et al., "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994-4.

[4] K. Ohkura et al., "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. ICSLP'92, pp. 369-

372, 1992.

[5] K. Ohtsuki et al., "Topic Extraction with Multiple Topic-words in Broadcast-news Speech," Proc. ICASSP'98, pp. I-329-332, 1998.

[6] T. Noreault, et al., "A Performance Evaluation of Similarity Measure; Document Term Weighting Schemes and Representations in a Boolean Environment," in R. N. Oddy ed. Information Retrieval Research, London, Butterworths, pp. 57-76, 1997.

[7] B. -H. Juang, "Automatic Speech Recognition: Problems, Progress & Prospects," IEEE Workshop on Neural Networks for Signal Processing, 1996.

[8] K. Ohtsuki et al., "Message-driven Speech Recognition and Topic-word Extraction," Proc. ICASSP'99, SP20-2, 1999.

[9] S. F. Chen, et al., "Topic Adaptation for Language Modeling using Unnormalized Exponential Models," Proc. ICASSP'98, pp. II-681-684, 1998.

[10] S. Miller, et al., "Statistical Language Processing using Hidden Understanding Models," Proc. DARPA Human Language Technology Workshop, pp. 278-282, 1994.

[11] R. Pieraccini, et al., "A Speech Understanding System based on Statistical Representation of Semantics," Proc. ICASSP'92, pp. I-193-196, 1992.

[12] R. Kuhn and R. De Mori, "A Cache-based Natural Language Model for Speech Recognition," IEEE Trans. PAMI-12, 6, pp. 570-583, 1990.

[13] Z. S. Harris, "Co-occurrence and Transformation in Linguistic Structure," Language, 33, pp. 283-340, 1957.