論文 / 著書情報
Article / Book Information

Title	Cluster-based modeling for ubiquitous speech recognition
Authors	Sadaoki Furui, Tomohisa Ichiba, Takahiro Shinozaki, Edward W.D.Whittaker, Koji Iwano
Citation	Interspeech2005, Vol. , No. , pp. 2865-2868,
Pub. date	2005, 9
Copyright	(c) 2005 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Cluster-based Modeling for Ubiquitous Speech Recognition

Sadaoki Furui, Tomohisa Ichiba, Takahiro Shinozaki, Edward W.D. Whittaker, and Koji Iwano

Department of Computer Science Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{furui, tichiba, staka, edw, iwano}@furui.cs.titech.ac.jp

Abstract

In order to realize speech recognition systems that can achieve high recognition accuracy for ubiquitous speech, it is crucial to make the systems flexible enough to cope with a large variability of spontaneous speech. This paper investigates two speech recognition methods that can adapt to speech variation using a large number of models trained based on clustering techniques; one automatically builds a model adapted to input speech using recognition hypotheses and clustered models, and the other directly uses clustered models in parallel. Both methods have been confirmed to be effective by evaluation experiments using presentation speech. Although the latter method needs a large amount of computation, it has an advantage in that it can be applied to online recognition, since it does not need recognition hypotheses. The former method can also be applied to online recognition, if the text of proceedings for the presentation can be used in place of recognition hypotheses.

1. Introduction

The demands of automatically transcribing ubiquitous speech for making spoken dialogue and speech document archiving systems are expected to rapidly increase in the near future. Although high recognition accuracy can be obtained for read speech, spontaneous speech is still very difficult to recognize. Since most of ubiquitous speech is spontaneous, it is crucial to improve the performance of spontaneous speech recognition. To achieve this goal, it is necessary to adapt both acoustic and language models to a wide range of variations, caused by speakers, topics, environments, contexts, etc. Adaptation methods can be classified into supervised and unsupervised methods. Since methods that need manual transcripts for supervision are impractical in many applications, it is desirable to develop unsupervised adaptation methods using the speech to be recognized as an information source for adaptation.

Although unsupervised adaptation methods have been successfully applied to the acoustic models for some time, relatively little work has been carried out in the area of language modeling (e.g. [1][2][3]). In general, the amount of data necessary for training language models is much larger than that needed for training acoustic models, that is, the language data space is much more sparse than the acoustic data space. Therefore, how to adapt language models using a limited amount of data is an important research issue in spontaneous speech recognition.

Among various unsupervised acoustic model adaptation techniques that have been so far investigated, methods using the MLLR (Maximum Likelihood Linear Regression) and cluster-based model selection techniques are considered to be easiest and most effective in adapting the models to various variations including the effects of speakers and environments [4][5]. Using gender-dependent models in parallel can be regarded as a special case of the cluster-based modeling. The methods using model selection are also effective for language model adaptation. We have investigated various language model adaptation methods based on model selection and class language models to improve the performance of spontaneous speech recognition [6][7].

This paper compares various adaptation methods and reports several new methods that we have recently proposed for recognizing spontaneous presentations in the *Corpus of Spontaneous Japanese* (CSJ) [8]. The paper is organized as follows. In Section 2 we describe a language model adaptation method using class language models. In Section 3 we describe a recognition method using many cluster models in parallel. Discussion and conclusion are given in Section 4.

2. Language model adaptation using class language models

2.1. Principles

We have investigated language models based on the combination of a general, word-based language model and multiple specialized, class-based language models, using linear interpolation as illustrated by the following [9]:

$$p(w \mid h) = \lambda_0 \cdot p_g(w \mid h) + \sum_{m=1}^{M} \lambda_m \cdot p_c(w \mid C(w), S) \cdot p_m(C(w) \mid C(h)) \qquad (1)$$

where w is the current word for which the probability is calculated, p_g is the general word n-gram language model built using data from the whole training corpus, h is the history, p_m is one of the M class n-gram models, λ_m is the weight assigned to each model such that $\Sigma \lambda_m = 1$ ($\lambda_m > 0$), p_c is the word-given-class unigram model, and S is the adaptation data source used to train p_c.

The word-class definitions $C(w)$ are trained on the whole training set, using the word clustering algorithm described by Kneser and Ney [10] to create $|C|$ different word classes where each word is a member of only one class. The class n-gram model is trained using a partition of the training data obtained by clustering similar presentations into M presentation clusters. Each cluster contains a certain number of presentations and each presentation is a member of a single cluster. The clustering method used is a bottom-up, agglomerative type of clustering based on a word co-occurrence metric. It was used in [2][11] and is based on [12]. The clustering is based on all the words from each presentation and the sequence of merge operations is preserved so that any desired number of clusters can be obtained. The interpolation weights $\lambda_0, \ldots, \lambda_m$ are

computed using the EM algorithm so as to maximize the likelihood of the adaptation data source S.

2.2. Adaptation using transcription hypothesis

We have investigated a method using the transcription hypothesis from the speech recognizer output as S for adapting p_c and $\{\lambda_m\}$. Multi-pass adaptation using the adaptation scheme has been performed. The first pass corresponds to recognition using the baseline n-gram language model with no adaptation. Each subsequent pass takes the transcription hypothesis from the previous pass for building the adapted model. It should be noted that, in our experiments, an entirely new recognition pass is performed for each pass rather than lattice rescoring being used. However, we believe it is unlikely that this has a significantly positive or negative effect on performance.

We perform recognition experiments using the Julius speech recognition engine version 3.3p3. In order to accommodate various combinations of word and word-class models, Julius was slightly modified such that language model probabilities could be obtained from an external library.

(a) Acoustic model
The acoustic features used for the experiments are 25 dimensional vectors consisting of 12 MFCCs, their delta as well as the delta log energy. All the models used are gender dependent triphone HMMs with 3000 shared states and 16 Gaussian mixtures. Cepstral mean subtraction (CMS) is also applied to each utterance. The acoustic models are trained using 2295 presentations (496 hours) by male speakers and 1154 presentations (218 hours) by female speakers in the CSJ.

(b) Baseline language model
The baseline language model is built from the transcribed content of about 2590 presentations in the CSJ, providing almost 7.5 million words of training data with a vocabulary size of 30678 words. Because there are generally no spaces between characters in written Japanese, the concept of a word boundary is not clearly defined. Thus, a word refers to a Japanese morpheme, extracted by a morphological analyzer.

All of the training data was used to build a baseline forward word bigram and a baseline reverse word trigram as needed by the Julius speech recognition engine. A variation of the smoothing technique developed by Kneser and Ney introduced in [13] is used with all language models.

(c) Experimental results
The test set, consisting of 30 presentations (20 academic presentations and 10 extemporaneous presentations, made by 20 male speakers and 10 female speakers), defined in the CSJ benchmark paper by Kawahara et al. [14] is used for testing. Among them, the set of 10 academic presentations made by male speakers is used as a development set.

The conditions of $M = 8$ and $|C| = 512$ are applied, since these conditions gave optimal performance on the development set [9]. The word error rate performance for two adaptation passes as well as the baseline (before adaptation) is given in Table 1. These results show that 9.1% relative improvement in word error rate can be obtained by the 2-pass adaptation method.

Table 2 shows the results for the 10 academic presentations, comparing the proposed unsupervised adaptation using up to three adaptation passes and supervised adaptation of the model when the correct transcription is used. These results show that improvements in performance are obtained using up to two unsupervised adaptation passes, and there is a significant difference with the result obtained by supervised adaptation.

Table 1: Word error rates (%) for multiple pass experiments, averaged over 30 presentations

Pass	Word error rate (%)
1 (Baseline)	26.5
2	24.5
3	24.1

Table 2: Comparison of unsupervised and supervised adaptation, averaged over 10 academic presentations

Adaptation method	Word error rate (%)
Baseline (1 pass)	26.7
Unsupervised (2 passes)	24.8
Unsupervised (3 passes)	24.3
Unsupervised (4 passes)	24.4
Supervised	23.6

2.3. Adaptation using proceedings text

The above-mentioned adaptation method has a problem in that it cannot be applied to online recognition, since it updates the language model using automatic transcripts of the whole presentation. On the other hand, proceedings describing the detail of presentations are often prepared and published before presentations, typically at conferences. Therefore, text in the proceedings can be used as S for language model adaptation as an alternative to automatic transcripts. This method has an advantage that it can be performed before presentation, and therefore it can be applied to online recognition.

Among the 30 presentations in the test set, 8 presentations given by 5 males and 3 female speakers are accompanied by proceedings. Therefore, they have been used to evaluate the adaptation method using the proceedings text. Table 3 compares the results of adaptation using proceedings text, and recognition hypotheses. These results show that by using the proceedings hypotheses, a slightly better result than that using the recognition hypotheses can be obtained.

Table 3: Comparison of adaptation by proceedings, or recognition hypothesis, averaged over 8 presentations

Adaptation method	Word error rate (%)
Baseline	24.8
By proceedings	22.6
By recognition hypotheses	22.8

2.4. Adaptation using both proceedings text and recognition hypotheses

A supplementary experiment using the 8 presentations has been performed to test the effectiveness of combining the two methods proposed in the previous subsections. The model adapted using proceedings text is used to obtain recognition hypotheses, and the model is further adapted by using the recognition hypotheses for the second run of recognition. As a result, 22.4% word error rate has been obtained, which is

slightly better than that obtained by either one of the adaptation methods.

3. Recognition method using many cluster models in parallel

Since computers are expected to become very small and cheap in the near future, it will soon become easy to use many computers (CPUs) in parallel, which each have different language and acoustic models, to recognize input speech. From this perspective, we have proposed combining cluster-based language and acoustic models based on the framework of a Massively Parallel Decoder (MPD) [15]. The MPD is a parallel decoder that has a large number of decoding units (DUs), in which each unit is assigned to each combination of element models, as shown in Fig. 1. An input speech utterance is sent to all the DUs and each DU independently processes the speech based on its language and acoustic model. The recognition hypotheses of the DUs as well as likelihood values are gathered by the integrator and a final output is produced. Since the system can be designed to run efficiently on a parallel computer, the turn around time is comparable to conventional decoders using a single model and processor.

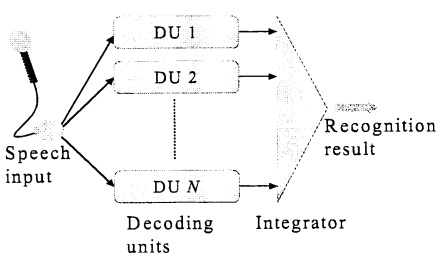

Fig 1: Architecture of the Massively Parallel Decoder (MPD).

3.1. Two types of cluster models

Using the presentation speeches in the CSJ, two types of cluster models have been investigated: one is based on presentation clustering and the other is based on utterance clustering. In the utterance clustering, all the utterances are independently clustered, irrespective of the presentation in which each utterance is included. 787 presentations given by male speakers with a total length of 186 hours and 2485 presentations containing 6.1 million words were used to build clusters for acoustic and language models, respectively.

3.2. Clustering for acoustic models

Clustering for acoustic models is conducted as follows:

Step 1: Randomly assign presentations/utterances to N clusters so that all the clusters have approximately the same number of presentations/utterances. Then make each cluster-based element model.

Step 2: Calculate likelihood of all the presentations/utterances for all the element models.

Step 3: Re-assign presentations/utterances to clusters based on their likelihood. The assignment is constrained so that all the clusters have the same number of presentations/ utterances.

Step 4: Make a cluster-based model.

Step 5: Return to Step 2 or terminate after sufficient number of iterations.

The likelihood values are calculated using triphone label files, and the number of iterations is set to 10. Based on the obtained definitions of the clusters, presentation/utterance cluster-based models are made by adapting the general model to each cluster using the MLLR adaptation method.

3.3. Clustering for language models

Clustering for language models is conducted using the same algorithm as that used for acoustic modeling, except that bigram perplexity is used as a measure instead of acoustic likelihood. Each component model is a word trigram which is trained by mixing the entire training set and the presentations/utterances in the cluster. This means that presentations/utterances in the cluster are weighted by duplication in the training set.

3.4. Experimental conditions

The first 10 academic presentations in the test set made by male speakers were used for evaluation. Utterances were extracted based on silence periods longer than 500 ms, and five minutes of utterances were extracted from each presentation. The subset therefore consists of 50 minutes of utterances, which corresponds to approximately half of the total duration of the ten presentations. Acoustic feature vectors had 38 elements comprising 12 MFCCs, their delta, delta-delta, delta log energy and delta-delta log energy. CMS was applied to each utterance.

A GRID system was used for the MPD. The baseline decoding system used the speaker-independent acoustic model (GAM) and the language model (GLM), whereas the MPDs used the cluster-based acoustic model and the cluster-based language model. The number of decoding units was a product of the number of elements of the cluster-based acoustic model and the cluster-based language model, where up to 400 decoding units were implemented. When presentation cluster-based models were used, the integrator selected recognition hypotheses throughout each presentation from one of the decoding units that maximized the total likelihood. On the other hand, when utterance cluster-based models were used, a hypothesis was selected independently for each utterance.

3.5. Experimental results

Table 4 shows the recognition results using the cluster-based acoustic models and language models. The baseline performance is different from that obtained in the previous experiment (Table 2), since various experimental conditions, especially acoustic features, are different. It can be seen that utterance clustering is consistently better than presentation clustering. The highest word error rate reduction rate of 7.7% was achieved using 10 acoustic models, and the reduction of 6.4% was achieved using 20 language models. By using 100 models as a result of the combination of 10 acoustic models and 10 language models, the word error rate reduction of 11.8% was obtained.

By applying unsupervised batch-type acoustic as well as language model adaptation to the MPD system having 100 models, a word error rate of 20.4% was obtained. The acoustic models were adapted by the MLLR method, whereas the language models were adapted by interpolating each

cluster-based language model with a cluster-independent word-class-based language model obtained using automatic transcripts of the whole presentation.

Table 4: Word error rate using the MPD based on cluster-based models

Cluster-based models	Number of clusters	Presentation clustering	Utterance clustering
Baseline	1	24.9	
Acoustic models only	5	24.0	23.7
	10	23.8	23.0
	20	23.8	23.2
Language models only	5	24.6	24.0
	10	24.7	23.6
	20	24.5	23.3
	40	24.3	23.6
Both acoustic and language models	5x5=25	23.7	22.9
	10x10=100	23.4	22.0
	20x20=400	23.5	22.1

Utterance-based cluster models gave significantly lower word error rates than presentation-based cluster models, probably for the following reasons. For language modeling, it is easier to find similar examples in the training set when the selection unit is shorter. For acoustic modeling, although the primary source of difference in acoustic characteristics is considered to be individuality, voice characteristics vary from utterance to utterance even in a single speaker.

4. Conclusions

This paper has proposed two methods using acoustic as well as language models made by clustering techniques for recognizing ubiquitous spontaneous speech having a wide range of variations. Recognition experiments using presentation speech have confirmed the effectiveness of the proposed methods. However, presentation is just one style of spontaneous speech, and our recent analysis has revealed that presentation speech is acoustically located in the middle between read speech and dialogue [8][16]. Our experiment on recognizing lecture speech in university classes has shown that they are acoustically as well as linguistically very different from presentations at conferences.

Therefore, we need to expand the range of target spontaneous speech to build a wider variety of acoustic and language models for recognizing ubiquitous spontaneous speech. For this purpose, it is necessary to build various corpora to cover the wide range of spontaneous speech. It will become crucially important to establish an efficient method of choosing the most appropriate model for each input speech according to its speaking style, environment, and context. Since it is very labor intensive to build large spontaneous speech corpora and there is a limit on the size of corpora that we can build, it is necessary to investigate how to efficiently sample a wide range of spontaneous speech.

Due to the rapid progress of computers, speech recognition methods using many models and CPUs are expected to become popular in the near future. By combining the two methods proposed in this paper, several new adaptive recognition methods are expected to emerge. Future works include improving clustering algorithms and investigating integration methods for recognition hypotheses from many decoding units.

5. Acknowledgments

This research has been supported by the "Spontaneous speech corpus and processing" project and the 21st Century Center-of-Excellence Program "Framework for Systematization and Application of Large-scale Knowledge Resources".

6. References

[1] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," Proc. ICASSP, Hong Kong, pp. 224-227, 2003.

[2] G. Moore and S. Young, "Class-based language model adaptation using mixtures of word-class weights," Proc. ICSLP, Beijing, pp. 512-515, 2000.

[3] T. R. Niesler and D. Willet, "Unsupervised language model adaptation for lecture speech transcription," Proc. ICSLP, Denver, pp. 1413-1416, 2002.

[4] Z. Zhang, S. Furui and K. Ohtsuki, "On-line incremental speaker adaptation for broadcast news transcription," Speech Communication, 37, pp. 271-281, 2002.

[5] Z. Zhang, T. Sugimura and S. Furui, "A tree-structured clestering method integrating noise and SNR for piecewise linear-transformation-based noise adaptation," Proc. ICASSP, pp. I-981-984, 2004.

[6] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, "Unsupervised class-based language model adaptation for spontaneous speech recognition," Proc. ICASSP, pp. 236-239, 2003.

[7] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, "Unsupervised language model adaptation using word classes for spontaneous speech recognition," Proc. SSPR, Tokyo, pp. 71-74, 2003.

[8] S. Furui, M. Nakamura, T. Ichiba and K. Iwano, "Analysis and recognition of spontaneous speech using *Corpus of Spontaneous Japanese*," Speech Communication, 2005. (to be published)

[9] L. Lussier, E. W. D. Whittaker and S. Furui, "Unsupervised language model adaptation methods for spontaneous speech," Proc. INTERSPEECH 2004 – ICSLP, Jeju, Spec4201o.4, 2004.

[10] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modeling," Proc. EUROSPEECH, pp 973-976, 1993.

[11] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models," Proc. ICSLP, pp. 236-239, 1996.

[12] S. Sekine, "Automatic sublanguage identification for a new text," Proc. Second Annual Workshop on Very Large Corpora, Kyoto, pp. 109-120, 1994.

[13] J. T. Goodman, "A bit of progress in language modeling," Technical Report, Microsoft Research, 2001.

[14] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," Proc. SSPR, Tokyo, pp. 135-138, 2003.

[15] T. Shinozaki and S. Furui, "Spontaneous speech recognition using a massively parallel decoder," Proc. INTERSPEECH 2004 – ICSLP, Jeju, ThA2001o.3, 2004.

[16] M. Nakamura, K. Iwano and S. Furui, "Analysis of cepstral features of Japanese spontaneous speech using Mahalanobis distance," Proc. Spring Acoustical Society of Japan Meeting, 2-1-14, 2005.