

論文 / 著書情報
Article / Book Information

Title	An Online Incremental Speaker Adaptation Method Using Speaker-Clustered Initial Models
Author	Zhipeng Zhang, Sadaoki Furui
Journal/Book name	ICSLP2000, Vol. 3, No. , pp. 694-697
発行日 / Issue date	2000, 10

AN ONLINE INCREMENTAL SPEAKER ADAPTATION METHOD USING SPEAKER-CLUSTERED INITIAL MODELS

Zhipeng Zhang and Sadaoki Furui

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

ABSTRACT

We previously proposed an incremental speaker adaptation method combined with automatic speaker-change detection for broadcast news transcription where speakers change frequently and each of them utters a series of several sentences. In this method, the speaker change is detected using speaker-independent and speaker-adaptive Gaussian mixture models (GMMs). Both phone HMMs and GMMs are incrementally adapted to each speaker by the combination of MLLR, MAP and VFS methods using speaker-independent (SI) models as initial models. This paper proposes its improvement in which an initial model for speaker adaptation is selected from a set of models made by speaker clustering. Either cluster-dependent phone HMMs or GMMs are used to calculate the likelihood for selecting the best initial model. In a broadcast news transcription task, the proposed method significantly reduces word error rate compared with the method using SI-HMM as an initial model. Online incremental speaker adaptation results show that word error rate is reduced by 11.6% relative to the baseline system with no speaker adaptation. The method using GMMs for cluster selection requires a significantly less number of computations than that using HMMs.

1. INTRODUCTION

Speaker-dependent (SD) speech recognition systems usually perform much better than speaker-independent (SI) systems. However, it is usually unrealistic to collect and use a sufficiently large set of training data for each speaker, especially for large-vocabulary continuous-speech recognition systems. Since speaker-adaptive (SA) systems can improve recognition performance for all speakers, many researchers have directed their attention to this area.

We previously proposed an incremental speaker adaptation method combined with automatic speaker-change detection for broadcast news transcription in which speakers change frequently and each of them utters a series of several sentences [1]. In this method, an adaptation technique combining MLLR [2],

MAP and VFS [3] is carried out using SI-HMM as an initial acoustical model for adaptation.

Speaker clustering scheme is one of the major techniques for speaker adaptation. In this method, a set of speakers is clustered and a cluster-dependent HMM is constructed for each cluster. In the recognition phase, a cluster which best fits input speech, that is a cluster which yields the largest likelihood is selected and an HMM corresponding to the selected cluster is used for decoding. The selected HMM can also be used as an initial model for speaker adaptation in place of SI-HMM. However, how to cluster speakers and how to select the best cluster model for speaker adaptation as well as for recognition are important research issues. How to reduce the amount of computation for likelihood calculation needed for selecting the best cluster by decoding each sentence using all the cluster-dependent HMMs is another important issue.

This paper investigates effectiveness of using a cluster-dependent HMM as an initial model for online incremental speaker adaptation and proposes a method to reduce the amount of calculation in cluster selection by using cluster-dependent GMMs instead of HMMs. These methods are tested in a broadcast news transcription task.

Section 2 describes methods of speaker clustering and cluster selection, and Section 3 reports broadcast news speech recognition experiments conducted to evaluate speaker adaptation methods using cluster-dependent HMMs or GMMs as initial models. The paper concludes with discussion and future research issues.

2. SPEAKER CLUSTERING

2.1 Clustering method

Since directly clustering speech data uttered by many speakers needs huge amount of computation, we cluster speaker-dependent models rather than directly clustering speech data. Speaker-dependent phone HMM (SD-HMM) is first built based on the Baum-Welch algorithm for each speaker

using all speech data uttered by the speaker. Since all the SD-HMMs have the same number of states, the distance between them can be defined according to the following equation:

$$d(i, j) = -\frac{1}{S * M} \sum_{s=1}^S \sum_{m=1}^M \{\log[b_{js}(\mu_{ism}) + \log[b_{is}(\mu_{jsm})]\}$$

where b_{is} represents the output probability distribution function of the m -th mixture in the S -th state, and S and M represent the number of states and the number of mixtures in each state respectively. In the distance calculation, only the mean values of distributions are taken into account, and the differences of covariance, transition probabilities and weights are neglected.

Based on a distance matrix consisting of the distances between the SD-HMMs, a clustering procedure originally proposed for the "SPLIT" speech recognition system [4] is carried out. This procedure has an advantage that any number of clusters can be made. As the number of clusters increases, the sum of likelihood values increases. The clustering is terminated automatically when the sum of likelihood or the number of clusters exceeds a pre-set threshold.

2.2 HMM-based cluster selection

Figure 1 shows a block diagram of a speech recognition system including the HMM-based speaker-cluster selection procedure. In this method, a test utterance is first decoded using the SI-HMM and a phone label sequence is produced. Phone HMMs in each cluster-dependent HMM (Cluster-HMM) are concatenated according to the label sequence and used for re-calculating the likelihood. The best Cluster-HMM yielding the maximum likelihood is selected and used to re-recognize the input utterance.

2.3 GMM-based cluster selection

The above-mentioned method has a disadvantage that it requires a large amount of computation in cluster selection, since it first recognizes an input utterance using SI-HMM and then re-calculate the likelihood using every Cluster-HMM. To alleviate this problem, we propose to use a GMM corresponding to each Cluster-HMM, in view of the success of using GMMs in text-independent speaker recognition [5]. For this purpose, SI-GMM is first made from all the data that were used in constructing the SI-HMM. Each cluster-dependent GMM (Cluster-GMM) is made from the same data that were used in constructing the corresponding Cluster-HMM. The number of mixtures in GMM is set to 64, based on the speaker recognition results. Figure 2 shows a block diagram of a speech recognition system including the speaker-cluster selection method based on the GMM-based measure. For each test utterance, the Clus-

ter-HMM corresponding to the GMM yielding the largest likelihood is selected and used for recognition.

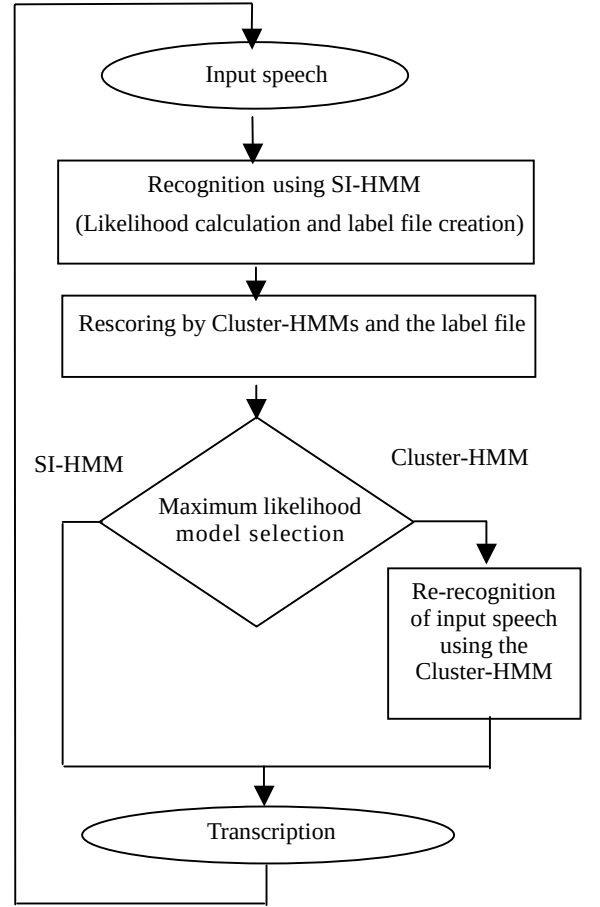


Figure 1: Speech recognition including HMM-based speaker cluster selection (SI: speaker independent)

3. EXPERIMENTS ON A JAPANESE BROADCAST NEWS TRANSCRIPTION SYSTEM

3.1 Language models

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, which comprised roughly 500k sentences and 22M words (morphemes). To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words (morphemes) using a morphological analyzer since Japanese sentences are written without spaces between words. Since many Japanese words

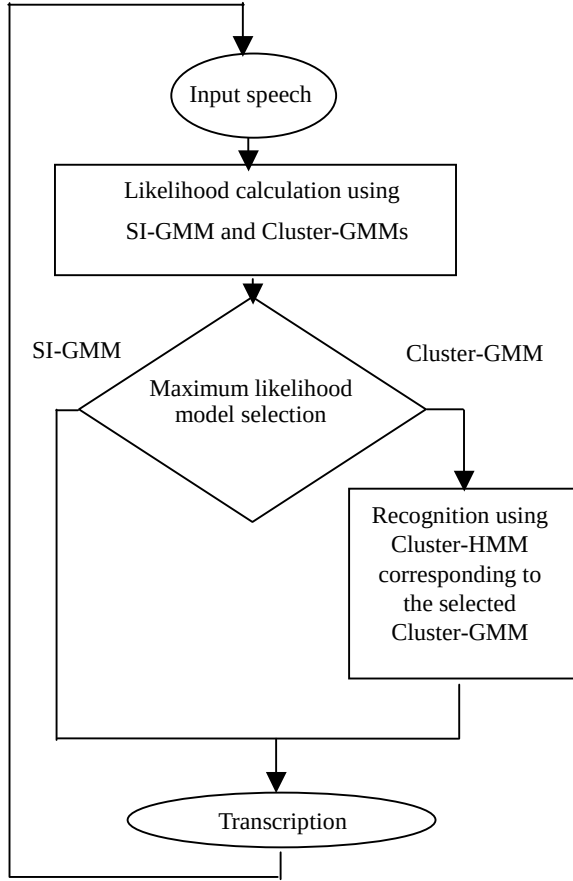


Figure 2: Speech recognition including GMM-based speaker cluster selection (SI: speaker independent)

have multiple readings and correct readings can only be decided according to context, we have constructed a language model in which a word with multiple readings is split into different language model entries according to its reading [6]. We also introduced filled-pause modeling into the language model. A word-frequency list was derived for the news manuscripts, and the 20k most frequently used words were selected as vocabulary words. The 20k vocabulary covered approximately 98% of the words in the broadcast-news manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using the Katz's back-off smoothing method.

3.2 Acoustic models

The feature vector extracted from speech consisted of 16 cepstral coefficients, normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector was 34. The 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 the tree-based clus-

tering. They were trained using the phonetically-balanced sentences and dialogues read by the 53 male and 56 female speakers, respectively. 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.

3.3 Evaluation data

Speech data consisting of 50 male and 50 female utterances with no background noise were extracted from TV broadcast news in July 1996 and used as evaluation utterances. The male and female sets respectively included utterances by five or six speakers. All utterances were manually segmented into sentences.

3.4 Speaker clustering and cluster-dependent HMM production

Speaker clustering was performed based on the clustering method described in Subsection 2.1 using the same utterances that were used in estimating the SI-HMM. The database consisted of phonetically-balanced sentences and dialogues as described in Subsection 3.2. After clustering, phone HMM for each cluster (Cluster-HMM) was made using all the utterances belonging to the cluster. Each Cluster-HMM was first estimated using the Baum-Welch algorithm and then interpolated with the SI-HMM.

3.5 Experimental results of speaker adaptation by cluster selection

Cluster-selection-based speaker adaptation experiment was performed. Figure 3 shows word error rates (WERs) for the female test set according to the number of clusters. The lowest error rate is obtained when the number of clusters is 4 for both male and female utterances.

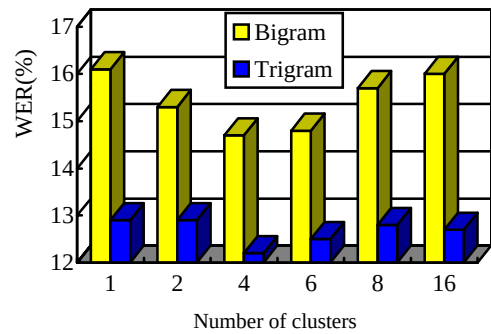


Figure 3: Word error rate for various numbers of clusters (Female test set)

Figure 4 shows experimental results in the 4-cluster case. This figure shows recognition results for the three conditions; no-adaptation (baseline), cluster selection using HMMs, and cluster selection using GMMs. The HMM- and GMM-based methods reduce the word error rate by 7.0% and 4.3%, respectively, averaged over male and female speakers, relative to the results for the baseline, in the case of the trigram language model.

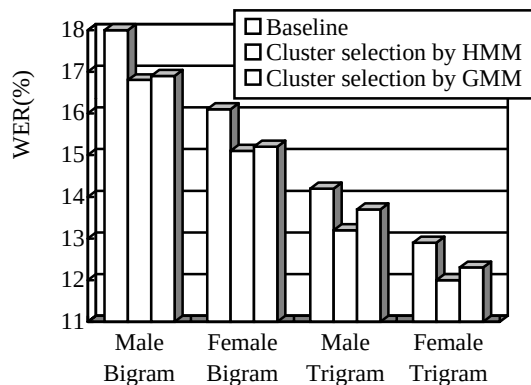


Figure 4: Word error rate by baseline and HMM/GMM-based cluster selection methods

3.6 Online incremental speaker adaptation using a cluster-dependent initial HMM selected by the GMM-based method

Online incremental speaker adaptation was carried out using the Cluster-HMMs as initial models. The best cluster for each input utterance is selected by the GMM-based cluster selection technique. Figure 5 shows recognition results for the three methods; no-adaptation (baseline), incremental speaker adaptation using SI-HMM as an initial model (1 cluster), and incremental speaker adaptation including cluster selection with 4 clusters. The 1-cluster and 4-cluster methods achieve 10.0% and 11.6% error rate reduction, respectively, averaged over male and female speakers, in comparison with the baseline, no-adaptation case using the trigram. The method proposed in this paper using a cluster-dependent initial HMM selected according to the GMM-based likelihood is significantly better than our previous method using a single cluster initial HMM, that is the speaker-independent HMM.

4. CONCLUSION

Speaker adaptation using initial models made by speaker clustering was investigated in this paper. Specifically the paper proposed a new model selection method that uses cluster-dependent GMMs instead of HMMs to reduce the amount

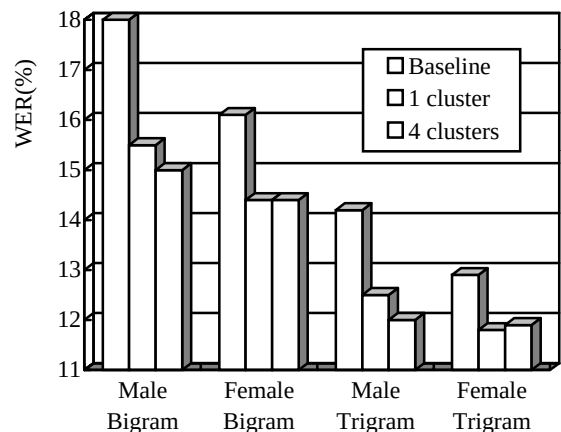


Figure 5: Word error rate for incremental adaptation using cluster-dependent initial HMMs selected by the GMM-based method

of computation. Use of this method achieved 11.6% word error rate reduction on the broadcast news transcription task.

Future research includes automatic audio signal segmentation into acoustically homogeneous periods and more general adaptation methods that can contend with not only speaker-to-speaker variability but also noise and channel variations to make recognition systems more robust against a wider range of mismatch conditions.

ACKNOWLEDGMENTS

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database. The authors also would like to thank Mr. Katsutoshi Ohtsuki at NTT Cyber Space Laboratories for his valuable discussions.

REFERENCES

- [1] Z.P. Zhang et al., Proc. ICASSP, pp. II-761-764, 2000
- [2] C. J. Leggetter et al., Computer Speech and Language, pp. 171-185, 1995.
- [3] K. Ohkura et al., Proc. ICSLP'92, pp. 369-372, 1992.
- [4] N. Sugamura et al., Proc. Speech Tech. Committee Meeting, ASJ, S82-64, 1982
- [5] T. Matsui et al., Proc. ICASSP, pp. II-157-160, 1992
- [6] K. Ohtsuki et al., Proc. Eurospeech'99, pp. 671-674, 1999