UNSUPERVISED SPEAKER ADAPTATION BASED ON SUFFICIENT HMM STATISTICS OF SELECTED SPEAKERS

*Shinichi Yoshizawa *, Akira Baba *, Kanako Matsunami **, Yuichiro Mera **, Miichi Yamada **, Kiyohiro Shikano ***

* Laboratories of Image Information Science and Technology
** Nara Institute of Science and Technology

ABSTRACT

This paper describes an efficient method for unsupervised speaker adaptation. This method is based on (1) selecting a subset of speakers who are acoustically close to a test speaker, and (2) calculating adapted model parameters according to the previously stored sufficient HMM statistics of the selected speakers' data. In this method, only a few unsupervised test speaker's data are required for the adaptation. Also, by using the sufficient HMM statistics of the selected speakers' data, a quick adaptation can be done. Compared with a pre-clustering method, the proposed method can obtain a more optimal speaker cluster because the clustering result is determined according to test speaker's data on-line. Experiment results show that the proposed method attains better improvement than MLLR [1] from the speaker independent model. Moreover the proposed method utilizes only one unsupervised sentence utterance, while MLLR usually utilizes more than ten supervised sentence utterances.

1. INTRODUCTION

Various kinds of speaker adaptation schemes have been proposed. A speaker dependent (-like) model is trained using a specific speaker's data or speakers' data close to the specific speaker. For using a lot of data for training, it takes a lot of time to make an acoustic model. Therefore, this type of model adaptation is difficult to be used in the on-line adaptation mode.

To solve the above problem, pre-clustering method has been proposed [2]. In this method, several speaker-dependent models are prepared before adaptation mode. It takes little time to obtain an adapted model, because the closest model for a test speaker is just selected on the adaptation mode. In this method, it is important to decide what kinds of speaker-dependent models are prepared.

MLLR [1] [6] [5] is a very popular scheme and it has been widely used. MLLR can obtain a large improvement of the recognition rate over a speaker-independent model. The combination of MLLR and the pre-clustering method [2] is also proposed. In general, to obtain a high improvement, a lot of adaptation data with the phoneme transcription are needed and it takes time for adaptation.

In this paper, a new adaptation method is proposed. This method is based on (1) selecting a subset of speakers who are acoustically close to a test speaker, and (2) calculating adapted model parameters according to the previously stored sufficient HMM statistics of the selected speakers' data. In this method, only a few unsupervised test speaker's data are necessary for the adaptation. Also, by using the sufficient HMM statistics of the selected speakers, a quick adaptation can be done. Compared with a pre-clustering method, the proposed method can obtain a more optimal cluster because the clustering result is determined according to the test speaker's data on-line. Experiment results show that the proposed method attains better improvement than those of MLLR [1].

2. BY SUFFICIENT STATISTICS SPEAKER ADAPTATION

The proposed method is described in Fig.1. This adaptation scheme consists of three steps. In the first step, a set of the parameters of sufficient HMM statistics for each speaker are calculated and pre-stored. In the second step, a subset of speakers who are acoustically close to the test speaker is selected using speaker models such as a Gaussian mixture model. The GMM speaker model is so simple that it can perform well even for a few test speaker's data without transcription. In the third step, an adapted acoustic model is calculated to combine the sufficient statistics from the speakers who are acoustically close to the test speaker.

In this paper, speech data are sampled at 16kHz and 16 bits. Twelfth-order mel-frequency cepstrum coefficients (MFCC) are calculated every 10 ms. The cepstrum differences (delta-MFCC) and delta-power are also used. Cepstrum mean normalization (CMN) is performed based on the whole utterance average.

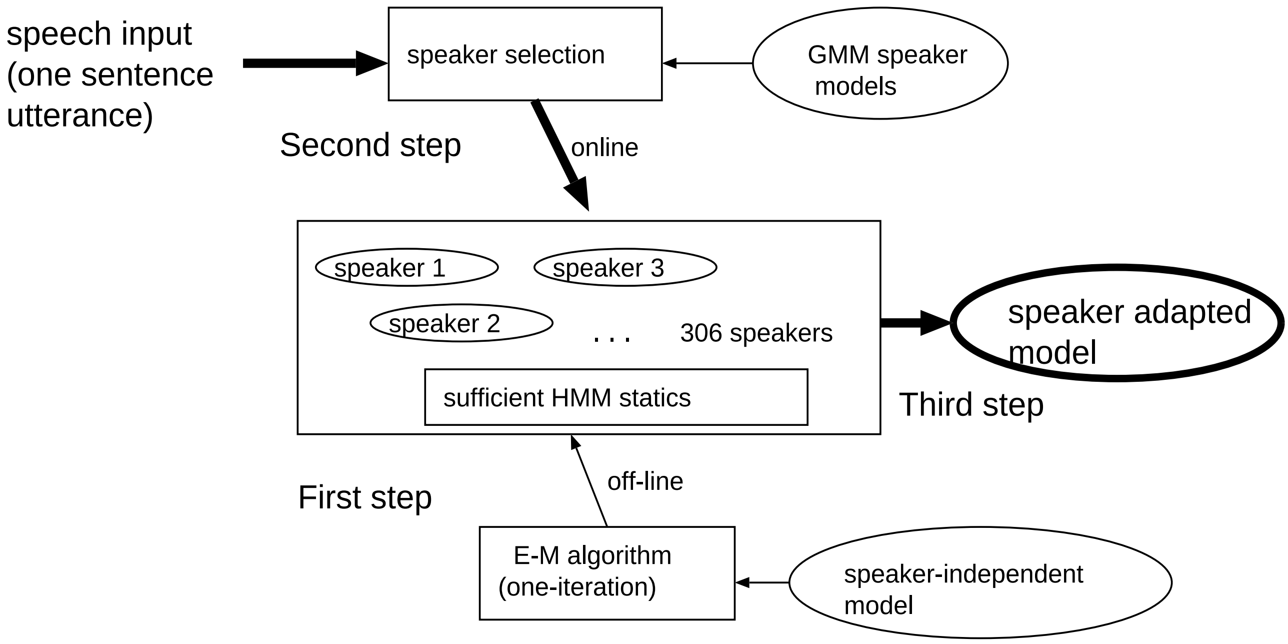

Figure 1: Blockdiagram of the proposed method based on speaker selection and sufficient HMM statistics

2.1. Calculating sufficient HMM statistics

Sufficient HMM statistics are the statistical parameters of the acoustic model, such as means, variances and E-M counts of hidden Markov models. The parameters are calculated for each speaker individually. The sufficient HMM statistics are estimated by one iteration of the E-M algorithm using each speaker's data and a speaker-independent HMM model.

2.2. Selecting a subset of speakers

In this section, a speaker selection using GMM is discussed. To obtain a good adapted model, it is important to select a subset of speakers who are acoustically close to a test speaker.

In this paper, for selecting a subset of speakers, speaker models consisting of the 64-Gaussian mixture model, which is a phone-independent one-state HMM, are used. As the distance between the test speaker's data and the other speakers' ones, the GMM acoustic likelihood for the adaptation data is used. These speaker model can perform well even for a few test speaker's data without phoneme transcription (in this paper, only one unsupervised sentence utterance is used for adaptation). Using this measure, the speakers are ordered according to the similarity to the test speaker. The top N-nearest speakers are selected as a subset of speakers for calculating the adapted acoustic model.

Compared with pre-clustering methods, the proposed method can obtain a more optimal cluster, which is called as a subset of speakers in this paper, because the subset is selected according to the test speaker's adaptation data and the cluster can be more adaptable than in the pre-clustering method.

2.3. Calculating adapted acoustic models

Performance evaluation is carried out using the Japanese dictation system Julius [4] with the 20k newspaper article language model.

Given some observation from a test speaker, a subset of speakers who are acoustically close to the test speaker is selected using the above procedure in section 2.2. In this section, we discuss how to make an acoustic model, which is adapted to a test speaker.

By introducing the concept of sufficient HMM statistics, it takes a little time to calculate an acoustic model in the adaptation procedure because these values can be calculated before adaptation off-line. In this method, instead of using database itself, the sufficient HMM statistics are used in the adaptation procedure. It requires almost no computation to create an adapted acoustic model from these parameters. This method has no inherent structure's limitation of transformation-based adaptation schemes such as MLLR [1] [5]. A speaker adapted acoustic model is calculated from the sufficient HMM statistics of the selected speakers using a statistical calculation method. This procedure is equivalent to the one-iteration of HMM training from the speaker-independent model.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental procedure is summarized below. Japanese speech corpus collected by Acoustical Society of Japan [3] is used in our experiments. This database consists of 306 speakers and each speaker uttered about 200 sentences.

As an acoustic model, two kinds of monophone models and Phonetic Tied Mixture (PTM) model [4] are used. PTM model is made from context-independent phone models with 64 mixture components per HMM state by assigning different mixture weights according to the shared states of triphones. PTM model can attain much better recognition rate than monophone models. PTM HMMs have totally 2500 states. Monophone HMMs of 43 phones have 3 states and each state has a mixture of 16 or 64 Gaussians.

46 speakers' data are used for testing data, which are not included in the training data. In the proposed method, an adapted model is calculated without using test speaker's sufficient statistics. In the proposed method, one unsupervised sentence adaptation utterance is used.

The baseline speaker-independent system shows the average word error rates of 18.5% (16 Gaussians), 13.5% (64 Gaussians) for the monophone models and 10.0% for the PTM model. The results of the standard MLLR adaptation [1] are described in Table1.

In Fig.2, the results for the proposed method are described. In this experiment, the effect of the number of selected speakers is investigated. From the figure, the minimum error rate of 14.9% (16 Gaussians), 10.9% (64 Gaussians) for the monophone models and 8.3% for the PTM model are attained. The proposed method attains better results than ones for MLLR by ten adaptation sentence utterances. As for the adaptation time, using the PTM model, the proposed method was roughly three times faster than MLLR by ten sentence adaptation utterances, and sixteen times faster than MLLR using the fifty sentence utterances in this experiment. These results are summarized in Table1 and Fig.2.

Table 1: Comparison with MLLR

method		proposed method	MLLR	
		unsupervised	supervised	
# of sentence utterances		1	10	50
word error rate	monophpone model (16 Gaussians)	14.9%	15.6%	13.8%
	monophpone model (64 Gaussians)	10.9%	12.6%	12.0%
	PTM (phonetic tied mixture model)	8.3%	9.0%	7.6%

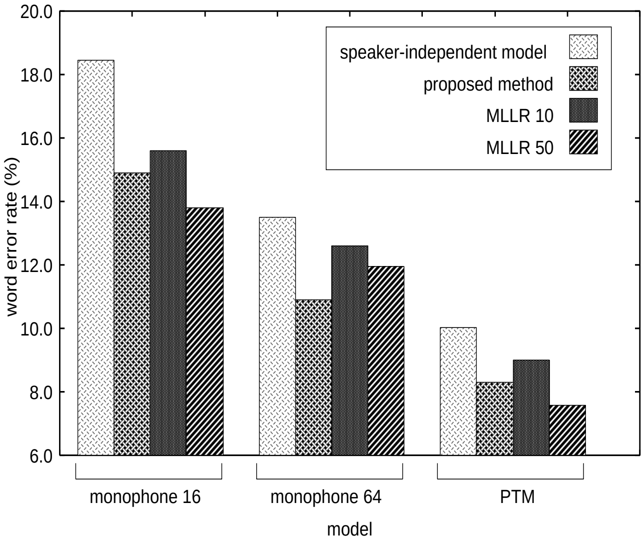

Figure 2: Comparison with various models

From the results in Table1 and Fig.2, the proposed method attains better recognition rates than the ones for MLLR by ten adaptation sentence utterances. The proposed method is especially efficient under the condition that only a small amount of adaptation data is available. MLLR needs more than ten sentence utterances for adaptation to attain the good recognition rate. And as for adaptation time, the proposed method is faster than MLLR for PTM. As the number of adaptation sentence utterances are increased, the difference of the adaptation time between the proposed method and MLLR becomes large and more critical.

In the proposed method, an unsupervised adaptation sentence utterance is used, but in MLLR more than ten supervised sentence utterances are required. Therefore, the proposed method is more useful to reduce a test speaker's effort.

As for the number of selected speakers, from the results in Fig.2, the optimum number are 20, 40 and 80 for the monophone with 16 Gaussians, the monohone with 64 Gaussians and PTM, respectively. The number of selected speakers becomes larger, as the model is more complicated. As for recognition rates, higher recognition rates are attained as the model is more complicated.

In Fig.4, the improvements of the accuracy for each speaker are shown, where the results are the best ones for PTM in which 80 speakers are selected for the adaptation. The horizontal axis notes test speakers who are sorted according to the word recognition accuracy of the pre-adaptation (speaker-independent) model. From the results, the low accuracy speakers are highly improved. The worst recognition

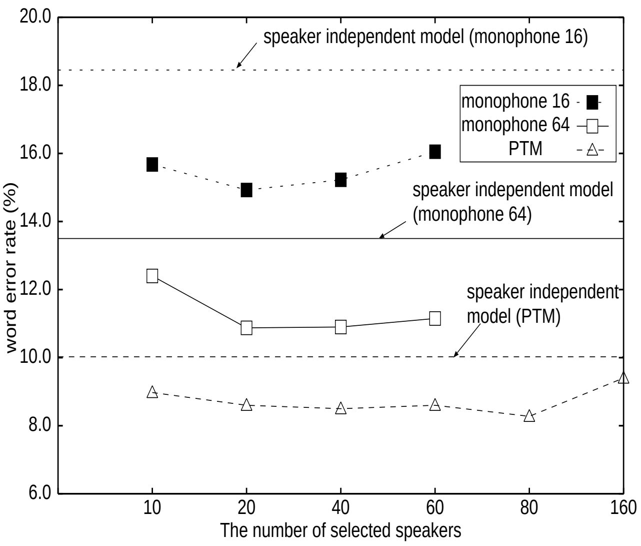

Figure 3: Word error rate for the proposed method

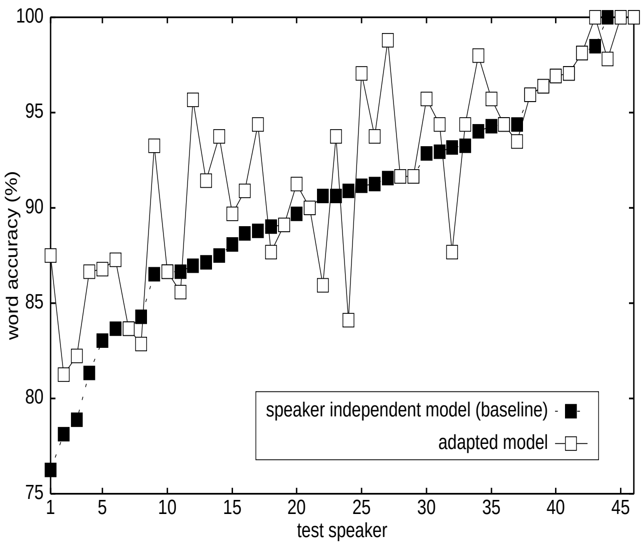

Figure 4: Improvement of word accuracy for each speaker using PTM model

rate is highly improved.

4. CONCLUSION

A new unsupervised adaptation method is proposed. This method is based on (1) selecting a subset of speakers who are acoustically close to a test speaker, and (2) calculating adapted model parameters according to the previously stored sufficient HMM statistics of the selected speaker's data. In this method, only a few unsupervised test speaker's data are necessary for the adaptation. By using the sufficient HMM statistics of the selected speaker's data, a quick adaptation can be done. Compared with a pre-clustering method, the proposed method can obtain a more optimal cluster because the clustering result is determined according to test speaker's data on-line. Experiment results show that the proposed method attains better improvement than those of MLLR. The proposed method is especially efficient under the condition that only a small amount of unsupervised adaptation data is available.

5. ACKNOWLEDGEMENT

This paper reports a part of work done by a project by NEDO (New Energy and Industrial Technology Department Organization).

6. REFERENCES

[1] C.J.Leggetter and C.Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, Vol. 9, pp. 171–185, 1995.

[2] Yuqing Gao, Mukund Padmanabhan, and Michael Picheny. SPEAKER ADAPTATION BASED ON PRE-CLUSTERING TRAINING SPEAKERS. *Proceedings of the Eurospeech*, pp. 2091–2094, 1997.

[3] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, and Shuichi Itahashi. JNAS:Japanese speech corpus for large vocabulary continuous speech recognition research. *The Journal of the Acoustical Society of Japan (E)*, Vol. 20, pp. 199–206, 1999.

[4] Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda, and Kiyohiro Shikano. A NEW PHONETIC TIED-MIXTURE MODEL FOR EFFICIENT DECODING. *Proceedings of the the ICASSP*, pp. 1269–1272, 2000.

[5] M.J.F.Gales and P.C.Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, Vol. 10, pp. 249–264, 1996.

[6] M.Padmanabhan, L.R.Bahal, D.Nahamoo, and M.A.Picheny. SPEAKER CLUSTERING AND TRANSFORMATION FOR SPEAKER ADAPTATION IN LARGE-VOCABULARY SPEECH RECOGNITION SYSTEM. *Proceedings of the the ICASSP*, pp. 701–704, 1995.