EVALUATION ON UNSUPERVISED SPEAKER ADAPTATION BASED ON SUFFICIENT HMM STATISTICS OF SELECTED SPEAKERS

*Shinichi Yoshizawa *, Akira Baba, Kanako Matsunami,*
Yuichirou Mera, Miichi Yamada, Akinobu Lee, Kiyohiro Shikano

* Laboratories of Image Information Science and Technology,Japan,
* Matsushita Electric Industrial Co.,Ltd.,Japan
yosizawa@crl.mei.co.jp

Abstract

This paper describes an efficient method of 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 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' 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. Experimental results show that the proposed method attains better improvement than MLLR from the speaker-independent model. The proposed method is evaluated in details and discussed.

1. INTRODUCTION

Pre-clustering method has been proposed [1]. In this method, several speaker-dependent models are prepared before adaptation mode. In this method, it is important to decide what kinds of speaker-dependent models are prepared.

MLLR [2] [3] 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 [1] 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 [6] and is evaluated in details. 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. Experimental results show that the proposed method attains better improvement than those of MLLR [3].

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.

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 model.

2.2. Selecting a subset of speakers

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. 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 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 for 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

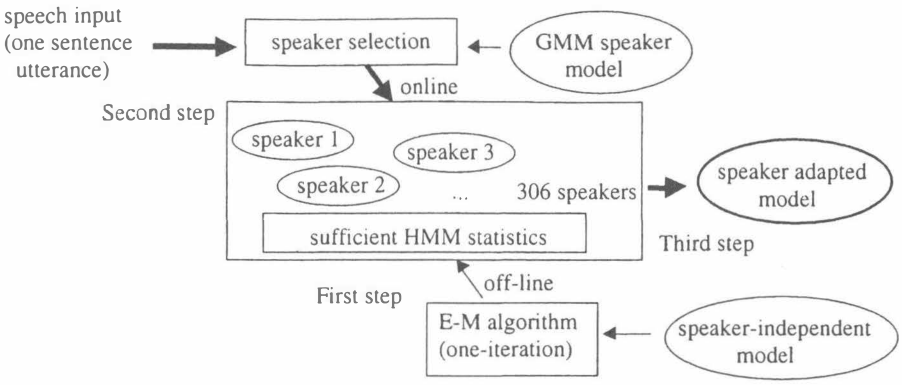

Figure 1: *Blockdiagram of the proposed method based on speaker selection and sufficient HMM statistics.*

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.

A speaker adapted acoustic model is calculated from the sufficient HMM statistics of the selected speakers using a statistical calculation method as follows:

$$\mu_i^{adp} = \frac{\sum_{j=1}^{N_{sel}} C^{j}_{mix}\mu_i^j}{\sum_{j=1}^{N_{sel}} C^j_{mix}} \qquad (i = 1, ..., N_{mix}) \qquad (1)$$

$$v_i^{adp} = \frac{\sum_{j=1}^{N_{sel}} C^j_{mix}(v_i^j + (\mu_i^j)^2)}{\sum_{j=1}^{N_{sel}} C^j_{mix}} - (\mu_i^{adp})^2 \qquad (i = 1, ..., N_{mix}) \qquad (2)$$

$$a^{adp}[i][j] = \frac{\sum_{k=1}^{N_{sel}} C^k_{state}[i][j]}{\sum_{j=1}^{N_{state}} \sum_{k=1}^{N_{sel}} C^k_{state}[i][j]} \qquad (i, j = 1, ..., N_{state}) \qquad (3)$$

where, μ_i^{adp}, v_i^{adp} $(i = 1, 2, ..., N_{mix})$, μ_i^j, v_i^j $(i = 1, ..., N_{mix}, j = 1, ..., N_{sel})$ are means and variances for the adapted model and for the selected speakers, respectively. $a^{adp}[i][j]$ $(i, j = 1, 2, ..., N_{state})$ is the transition probability from state i to state j. N_{mix} and N_{state} are the number of Gaussians and of states, respectively. C^j_{mix} $(j = 1, ..., N_{sel})$ and $C^k_{state}[i][j]$ $(k = 1, ..., N_{sel}, i, j = 1, ..., N_{state})$ are E-M counts for Gaussians and for states transition, respectively.

This procedure is equivalent to the one-iteration of HMM training from the speaker-independent model.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Japanese speech corpus collected by Acoustical Society of Japan [5] is used in our experiments. This database consists of 306 speakers and each speaker uttered about 200 sentences.

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

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 for speaker-independent models. 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.

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

In the experiment, a little different parameters are used from ones in the paper [6] and better results are attained.

3.1. Comparison with MLLR

The baseline speaker-independent system shows the average word error rates of 18.1% (16 Gaussians), 13.6% (64 Gaussians) for the monophone models and 8.9% for the PTM model. The results of the standard MLLR adaptation [3] are described in Table 1 and Fig.2.

In Table1 and Fig.2, the results for the proposed method are described. From the results, the proposed method attains smaller word error rates than the ones for MLLR by ten adaptaion sentence utterances. By the monophone with 64 Gaussians or PTM as an acoustic model, the proposed method attains smaller word error rates than the ones for MLLR by fifty adaptaion sentence utterances. MLLR needs more than ten sentence utterances for adaptation to attain a good recognition rate. As for the adaptation time (except the time to utter adaptation sentences), 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.

3.2. The number of selected speakers

The effect of the number of selected speakers is investigated. From the results in Fig.3, the minimum error rate of 14.7% (16 Gaussians), 10.8% (64 Gaussians) for the monophone models and 6.6% for the PTM model are attained. The optimum num-

Table 1: *Comparison with MLLR.*

method		proposed method	MLLR		speaker-independent model
		unsupervised	supervised		
# of sentence utterances		1	10	50	---
word error rate	monophone model (16 Gaussians)	14.7%	15.9%	12.8%	18.1%
	monophone model (64 Gaussians)	10.8%	12.8%	11.6%	13.6%
	PTM (Phonetic Tied Mixture)	6.6%	7.5%	7.0%	8.9%

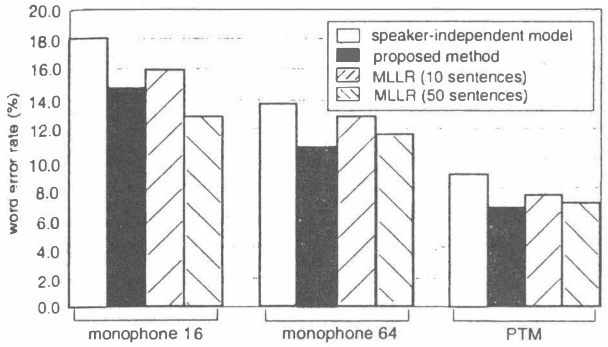

Figure 2: *Comparison with MLLR.*

ber are 20, 20 and 40 for the monophone with 16 Gaussians, the monophone with 64 Gaussians and PTM, respectively. The number of selected speakers becomes larger, as the model is more complicated.

3.3. Improvement of word acuuracy for each speaker

The improvements of the word accuracy for each speaker are shown in Fig.4,5, 6 and 7.

In Fig.4, the best result for PTM, in which 40 speakers are selected for the adaptation, is shown. The horizontal axis notes test speakers who are sorted according to the word recognition accuracy of the pre-adaptation (speaker-independent) model. From the result, the low accuracy speakers are highly improved. The worst recognition rate is highly improved.

In all above experiments, 4 or 5 sentences for each speaker are used. To evaluate the experiment in Fig.4 statistically, the sentences for each test speakers are increased into about 100 sentences. Fig.5 shows the result for PTM (40 speakers are selected for the adaptation). From the result, the word recognition accuracy for almost all speakers are improved. The results for the monophone models are shown in Fig.6 and 7. From the results, the word recognition accuracy for all speakers are improved. Total word recognition accuracy becomes lower because of using many sentences for evaluation and a lot of unknown words are included.

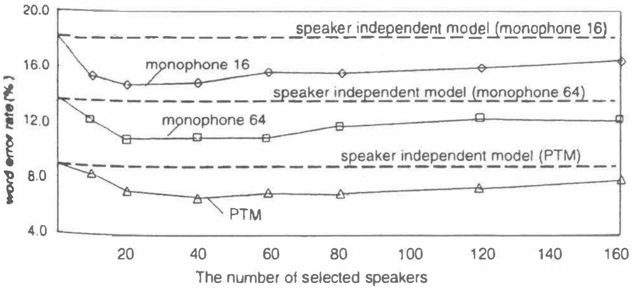

Figure 3: *Word error rate for the number of selected speakers.*

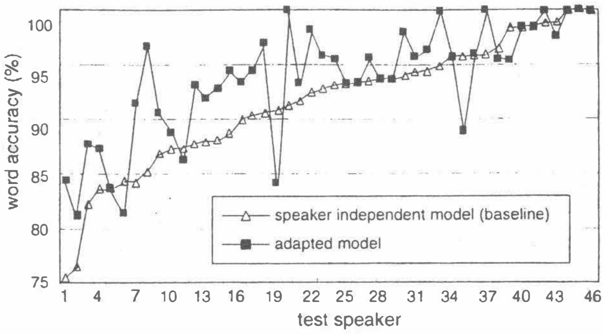

Figure 4: *Improvement of word accuracy for each speaker using PTM model (4 or 5 sentences for each speaker).*

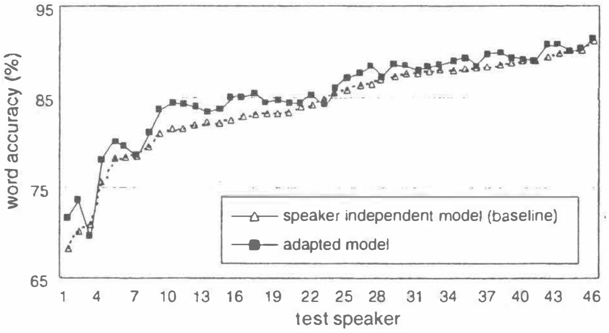

Figure 5: *Improvement of word accuracy for each speaker using PTM model (about 100 sentences for each speaker).*

3.4. Various methods for selecting speakers

Three different types of methods for selecting speakers are considered: (1) the number of selected speakers is decided for each test speaker, (2) the number of selected speakers is decided for each phonemic HMM (phoneme optimum method), and (3) the GMM speaker model is changed into one which is made from the frames having larger power than the average (large power speaker model method).

In Fig.8, word recognition accuracy for each speaker is shown. About 100 sentences is used for each test speaker. The monophone model with 16 Gaussians is used. From the results, almost all test speakers have the best results by selecting about 20 speakers for adaptation. If the best number of selected speakers for each speaker can be determined, a large improvement of recognition rate can be attained.

In Table2, the results of two other methods are describled (phoneme optimun method, and large power speaker model method). 4 or 5 sentences is used for each test speaker. In Table2, the original is a method in section2.2. From the results, these two methods attain the word error rates similar to the original one.

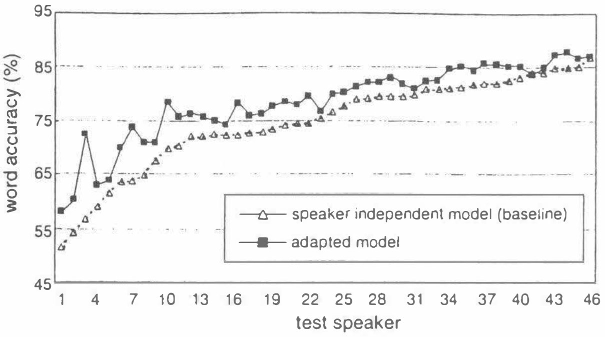

Figure 6: *Improvement of word accuracy for each speaker using monophone 16 Gaussians model (about 100 sentences for each speaker).*

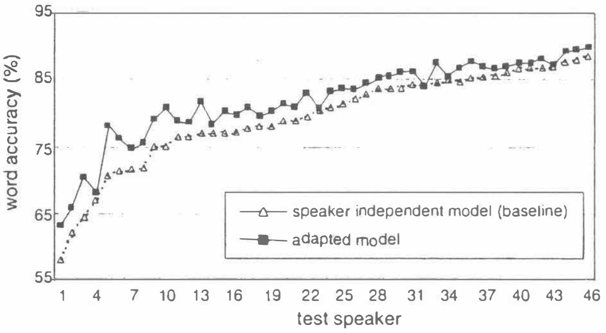

Figure 7: *Improvement of word accuracy for each speaker using monophone 64 Gaussians model (about 100 sentences for each speaker).*

4. CONCLUSION

A new adaptation method is proposed. 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 speakers' data, a quick adaptation can be done. Experimental results show that the proposed method attains better improvement than those of MLLR and it is evaluated in detail and discussed.

5. ACKNOWLEDGMENT

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

6. References

[1] Yuqing Gao, Mukund Padmanabhan and Michael Picheny, "SPEAKER ADAPTATION BASED ON PRE-CLUSTERING TRAINING SPEAKERS", Proceedings of the Eurospeech, pp.2091–2094, 1999.

[2] 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 ICASSP, pp.701–704, 1995.

[3] 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.

[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] 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.

[6] Shinichi Yoshizawa, Akira Baba, Kanako Matsunami, Yuichiro Mera, Miichi Yamada, Kiyohiro Shikano, "Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers", Proceedings of the ICASSP, 2001.

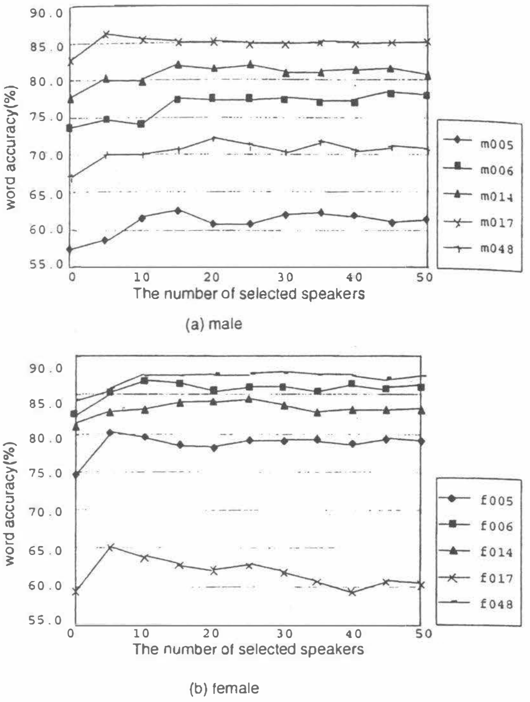

Figure 8: *Word recognition accuracy for each speaker (monophone model with 16 Gaussians).*

Table 2: *Various methods for selecting speakers.*

method		proposed method			MLLR	
		original	phoneme optimum	large power speaker mode		
# of sentence utterances		1	1	1	10	50
word error rate	monophone model (16 Gaussians)	14.7%	14.4%	14.9%	15.9%	12.8%
	monophone model (64 Gaussians)	10.8%	11.0%	11.1%	12.8%	11.6%