SPEAKER ADAPTATION BASED ON TRANSFER VECTOR FIELD SMOOTHING USING MAXIMUM A POSTERIORI PROBABILITY ESTIMATION

Masahiro Tonomura *Tetsuo Kosaka* *Shoichi Matsunaga*

ATR Interpreting Telecommunications Research Labs.,
2–2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

ABSTRACT

This paper proposes a novel speech adaptation algorithm that enables adaptation even with a small amount of speech data. This is a unified algorithm of two efficient conventional speaker adaptation techniques, which are maximum a posteriori (MAP) estimation and transfer vector field smoothing (VFS). This algorithm is designed to avoid the weaknesses of both MAP and VFS. A higher phoneme recognition performance was obtained by using this algorithm than with individual methods, showing the superiority of the proposed algorithm. The phoneme recognition error rate was reduced from 22.0% to 19.1% using this algorithm for a speaker-independent model with seven adaptation phrases. Furthermore, a priori knowledge concerning speaker characteristics was obtained for this algorithm by generating an initial HMM with the speech of a selected speaker cluster based on speaker similarity. The adaptation using this initial model reduced the phoneme recognition error rate from 22.0% to 17.7%.

1. INTRODUCTION

Recently, speaker adaptation methods have been studied for continuous mixture density HMM (CDHMM) based speech recognition systems. The key point of the speaker adaptation is how to estimate the model parameters with a small amount of adaptation data in order to obtain a high recognition performance. In addressing this problem, it is essential to consider compensating for information lacking from the adaptation data by using information from an initial model as a priori knowledge. However, the speaker adaptation then has the following problems.

- How to obtain a good initial model.
- How to combine a priori knowledge with a posteriori knowledge.
- How to interpolate the untrained parameters without training data.

An initial HMM must compensate for information lacking from the adaptation data when training data is sparse. Usually, a well-trained speaker-independent (SI) model that includes the information of a large population of speakers, and is used as an initial HMM, includes not only effective information but also inadequate information for an adapted speaker. However a good initial model is one that can reduce inadequate information and extract useful information for a target speaker selectively. Gender-dependent models are one example of good models and their effectiveness has been reported[1]. In this paper, in order to get a priori knowledge concerning more detailed speaker characteristics, a speaker-clustering (SC) model is used as an initial model. The SC model is generated by using the speech of a selected speaker cluster based on speaker similarity. The cluster selection is performed with a tree-structured speaker clustering technique[2] which selects the size and members of the cluster simultaneously using an utterance of a target speaker.

For the second problem, maximum a posteriori (MAP) estimation gives a good solution[1][3][4]. MAP estimation uses information from an initial model as a priori knowledge to compliment the training data. This a priori knowledge is statistically combined with a posteriori knowledge derived from the training data based on the amount of training data. In this method, however, some parameters having no training data are not trained, so interpolating these parameters is required.

The remaining problem is about untrained or insufficiently trained parameters without training data. To deal with this problem, transfer vector field smoothing (VFS) has been proposed[5]. This technique is based on the assumption that one speaker's acoustic feature space is transferable to another's continuously. Under this assumption, the VFS deals with the adaptation of mean vectors of Gaussian distributions in CDHMMs.

0-7803-2431-5/95 $4.00 © 1995 IEEE

The VFS algorithm consists of three steps, i.e. 1) Estimation of transfer vectors, 2) Interpolation, and 3) Smoothing. The transfer vectors imply the difference of the mean vectors between the initial model and the target model. Through the transfer vectors, the VFS also uses the information (i.e. mean of Gaussian distributions) of the initial model. In step 1), the transfer vectors are obtained by Maximum Likelihood (ML) estimation for the training data. In step 2), the untrained mean vectors having no training data are interpolated with the trained transfer vectors. In step 3), each transfer vector is smoothed with the other transfer vectors. Usually, the amounts of training data of the individual transfer vectors are not in balance when the training data is sparse. For this reason, the reliability of each estimated transfer vector is also different from the others. For each step of the VFS, therefore, consideration on the reliability of transfer vectors is required.

In this paper, the usage of an SC model as an initial model of speaker adaptation is proposed. In addition, a unified algorithm of the above two techniques (MAP-VFS) is proposed to solve the problems of individual methods and to achieve effective adaptation by using information from an initial model.

This paper is organized as follows. Section 2 gives details of the MAP-VFS algorithm. In Section 3, an application of this algorithm to an SI model and the subsequent evaluation are reported. Furthermore, to evaluate the effectiveness of an initial model, an SC model generated by using information of similar speakers to the adapted speaker is used as an initial model. The evaluation of the MAP-VFS algorithm with this initial model is shown.

2. MAP-VFS ALGORITHM

The MAP-VFS algorithm proposed here integrates MAP and VFS by the following three steps. First, the transfer vectors are estimated by MAP estimation. Then, interpolation and smoothing are performed using the transfer vectors obtained by the MAP estimation. Usually, when adapting parameters of CDHMMs using a small amount of training data, it is more effective to adapt the mean vectors of Gaussian distributions only than to adapt with the variance, transition probability and mixture gain for each Gaussian mixture. Therefore, in this paper, only the mean of each Gaussian distribution is adapted; the other parameters are not adapted.

2.1. Estimation of Transfer Vectors with MAP

The mean vectors of the Gaussian distributions of the initial model are recalculated through concatenation training. Let μ_p^I and μ_p^R be the p-th mean vector of the Gaussian distribution of the initial model and the retrained model, respectively. Then, the transfer vector for the p-th mean vector v_p is the difference between μ_p^R and μ_p^I.

$$v_p = \mu_p^R - \mu_p^I \tag{1}$$

where $p \in K_1$ (K_1 is the Gaussian distribution set having the training data). In this adaptive training, each mean vector μ_p^I of the initial model is used as the mean of an a priori distribution. Then, the MAP estimates of the mean μ_p^R are solved by[6]

$$\mu_p^R = \frac{n}{n+\tau} m_p + \frac{\tau}{n+\tau} \mu_p^I \tag{2}$$

where m_p is the sample mean of the p-th Gaussian distribution and also the Maximum Likelihood estimate, n is the total number of training samples observed for the corresponding Gaussian mixture component, and τ indicates the relative balance between the a prior knowledge and empirical data. Substituting this into Eq.(1), the transfer vector v_p is given by

$$v_p = \mu_p^R - \mu_p^I = \frac{n}{n+\tau}(m_p - \mu_p^I) \tag{3}$$

Then, the transfer vector v_p which is obtained by MAP estimation is represented as follows with the transfer vector v_p^{ML} ($= m_p - \mu_p^I$) which is estimated by Maximum Likelihood (ML) estimation.

$$v_p = \frac{n}{n+\tau} v_p^{ML} \tag{4}$$

This equation shows that the transfer vector with MAP estimation is obtained by modifying the transfer vector v_p^{ML} with ML estimation by the weighting coefficient which is a function that depends on the amount of training data n.

2.2. Interpolation of transfer vectors

Because of the limited amount of training data, not all mean vectors can be trained. These untrained mean vectors are characterized by μ_q^I ($q \in K_2$: K_2 is the set of Gaussian distributions having no training data). The transfer vector v_q of μ_q^I is interpolated according to the following equation using the trained transfer vectors v_p.

$$v_q = \sum_{k \in N(q)} \lambda_{q,k} v_k \Big/ \sum_{k \in N(q)} \lambda_{q,k} \tag{5}$$

where $N(q)$ is the set of k-nearest neighbors to μ_q^I. The $\lambda_{q,k}$ is the weighting coefficient that depends on the distance between μ_q^I and μ_k^I. Then, μ_q^I is transferred to μ_q^R using the interpolated transfer vector v_q.

$$\mu_q^R = \mu_q^I + v_q \tag{6}$$

2.3. Smoothing of transfer vectors

For all of the trained transfer vectors $v_p (p \in K_1)$, a smoothing operation is performed with Eq.(7). This smoothing operation assumes that all of the transfer vectors are combined with the restriction of continuity. In other words, one speaker's acoustic feature space is assumed to be transferable to another's continuously.

$$v_p^S = \sum_{k \in N(p)} \lambda_{p,k} v_k \ / \sum_{k \in N(p)} \lambda_{p,k} \qquad (7)$$

Then, the mean values of Gaussian distributions can be modified using the smoothed transfer vectors by Eq.(8).

$$\mu_p^S = \mu_p^I + v_p^S \qquad (8)$$

In this paper, six nearest neighbors are used for both interpolation and smoothing. The weighting coefficient $\lambda_{a,b}$ is calculated by Eq.(9).

$$\lambda_{a,b} = \exp\left(\frac{-d_{a,b}}{f}\right) \qquad (9)$$

where $d_{a,b}$ is the distance between μ_a^I and μ_b^I, and f is a weight control parameter.

3. EXPERIMENTS

3.1. Experimental conditions

We examined the proposed MAP-VFS speaker adaptation algorithm on Japanese 26-phoneme recognition. The experimental conditions are listed in Table 1. A 200-state Hidden Markov Network (HMnet)[7] is used to allow us to efficiently share system parameters in phoneme context-dependent HMMs. The number of mixture components per state is five for the SI model, and five or less for the SC model. Diagonal covariance matrices are used. Considering the dependency on training data for the speaker adaptation performance, the experiment is repeated five times for different training data selections.

3.2. Adaptation for Speaker-Independent Model

Speaker adaptation is performed for a SI model generated by using the speech of 285 speakers.

Figure 1 shows the phoneme recognition rate of the following four speaker adaptation methods: 1) re-training by maximum likelihood estimation (ML), 2) re-training by MAP estimation (MAP), 3) VFS with ML estimation (VFS), and 4) VFS with MAP estimation (MAP-VFS) which is the proposed method. MAP-VFS achieves a higher performance than MAP and VFS, each of which still shows a higher performance than ML. This shows that the proposed MAP-VFS method successfully unites these two methods.

Table 1: Experimental Conditions

Analysis conditions	
Sampling frequency 12kHz Hamming window 20ms, Frame period 5ms	
Analysis	
16-order LPC-Cepstrum + 16-order ΔLPC-Cepstrum + log power + Δlog power	
Training data	
146 males + 139 females (50 Japanese sentences per person)	
Adaptation/Recognition data	
Speakers	4 males (MAU,MMY,MSH,MTM) 3 females (FAF,FMS,FYM)
Adaptation	Random sampled N phrases ($N = 1,3,5,7$) from 256 Japanese phrases (SB1 task)
Recognition	279 Japanese phrases (SB3 task)

Table 2 shows that MAP-VFS achieves stable adaptation in proportion to the increase in the number of adaptation phrases for all speakers. Compared with the SI results, the phoneme error reduction is 5.3 ~ 19.9% (average of all speakers: 13.2%) by performing the speaker adaptation with the MAP-VFS method using seven adaptation phrases.

3.3. Adaptation for Speaker-Clustering model

To get a priori knowledge concerning speaker characteristics, speaker adaptation is performed for the SC model which is generated by using the speech of a speaker cluster automatically selected from out of 285 speakers with a tree-structured speaker clustering algorithm.

Table 3 shows that the phoneme error reduction is 6.8 ~ 34.4% (average of all speakers: 19.5%) over the SI results with seven adaptive phrases. Comparing this with the result of the previous section (13.2%) which is the experimental result directly adapted from the SI model with the MAP-VFS method, it is clear that the a priori knowledge from the SC model is effectively combined with the MAP-VFS method. Looking at the results for the individual speakers, those who had lower performances with the SI model (FYM, MSH etc.), showed remarkable performance improvements. This confirms the effectiveness of adaptation using the a priori knowledge of the SC model in the MAP-VFS method. On the other hand, for speakers who had higher performances with the SI model (MTM, MAU etc.), the SC model had little effectiveness. Overall the SC model with MAP-VFS, achieved low recognition error rates (less than 21.4%) using only seven adaptive phrases for all evaluated speakers. Because obtaining a good performance for all speakers is an important factor in speaker-independent speech recognition based on speaker adaptation, this result confirms the usefulness of the combined approach.

4. CONCLUSIONS AND FUTURE WORK

In this paper, a unified MAP-VFS algorithm has been proposed. We have examined that this unified method allows individual methods, i.e. MAP and VFS, to compliment each other's weaknesses. The effectiveness of this method has been shown in Japanese phoneme recognition experiments and a higher performance has been obtained compared with that of individual methods. Furthermore, considering the importance of a priori knowledge for the MAP-VFS method, we have adopted a tree-structured speaker clustering method to obtain a good initial model. Using this initial model, a higher phoneme recognition performance was obtained than that possible with an SI model as the initial model. As a result, the phoneme recognition error rate was reduced from 22.0% to 17.7% only with seven adaptation phrases, which corresponds to utterances of about 6.3 seconds. We plan to apply this method to on-line adaptation and unsupervised adaptation.

5. REFERENCES

[1] J.-L. Gauvain and C.-H. Lee: "Speaker adaptation based on map estimation of hmm parameters," Proc. of ICASSP 93, pp. 558-561(1993)

[2] T. Kosaka and S. Sagayama: "Tree-Structured Speaker Clustering For Fast Speaker Adaptation," Proc. of ICASSP 94, pp. 245-248(1994)

[3] C.-H. Lee, C.-H. Lin and B.-H.Juang: "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," IEEE Trans. on Signal Processing, Vol. 39, No. 4 (1991)

[4] J.-L. Gauvain and C.-H. Lee: "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2 (1994)

[5] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of ICSLP 92, pp. 369-372(1992)

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973

[7] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP 92, pp. 573-576(1992)

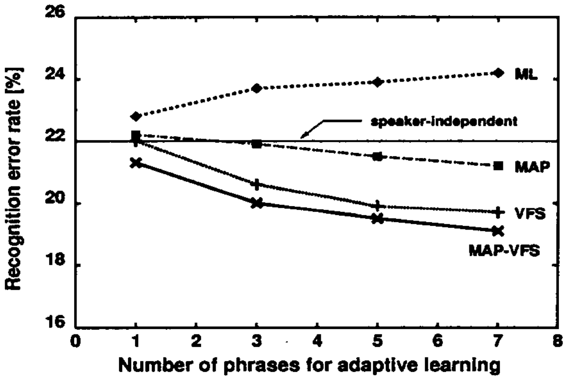

Figure 1: Phoneme recognition error rates for four speaker adaptation methods (average of 4 males and 3 females)

Table 2: Results of speaker adaptation from SI model using MAP-VFS method(%)

speaker	SI model	total training phrase numbers			
		1	3	5	7
MAU	18.2	18.3	17.5	16.7	16.4
MMY	19.0	19.2	18.4	18.0	18.0
MSH	24.1	23.7	22.1	21.6	21.0
MTM	17.5	16.6	15.9	15.6	15.0
FAF	21.5	20.4	18.9	17.9	18.1
FMS	21.1	20.5	19.6	19.3	19.0
FYM	32.6	30.4	27.6	27.1	26.1
average	22.0	21.3	20.0	19.5	19.1

Table 3: Results of speaker adaptation from SC model using MAP-VFS method (%) (values in parentheses are results for the SC model)

speaker	SI model	total training phrase numbers			
		1	3	5	7
MAU	18.2	21.6 (21.0)	16.9 (18.1)	16.2 (18.1)	15.9 (18.1)
MMY	19.0	21.2 (20.1)	18.2 (18.5)	17.7 (18.2)	17.7 (18.5)
MSH	24.1	23.8 (23.4)	20.2 (21.0)	19.0 (21.0)	18.5 (21.0)
MTM	17.5	19.1 (19.4)	16.1 (17.3)	15.8 (17.3)	15.4 (17.1)
FAF	21.5	18.6 (18.8)	17.7 (19.0)	16.9 (19.0)	17.1 (19.0)
FMS	21.1	18.6 (19.3)	17.8 (18.8)	17.4 (19.3)	17.8 (19.7)
FYM	32.6	25.8 (24.4)	22.7 (22.1)	21.6 (22.5)	21.4 (22.1)
average	22.0	21.2 (20.9)	18.5 (19.3)	17.8 (19.3)	17.7 (19.4)