VECTOR-FIELD-SMOOTHED BAYESIAN LEARNING FOR INCREMENTAL SPEAKER ADAPTATION

Jun-ichi TAKAHASHI *Shigeki SAGAYAMA*

NTT Human Interface Laboratories
1-2356 Take, Yokosuka, Kanagawa 238-03 Japan

ABSTRACT

This paper presents a fast and incremental speaker adaptation method called MAP/VFS, which combines Maximum a posteriori (MAP) estimation, or in other words Bayesian learning, with Vector Field Smoothing (VFS). The point is that MAP is an intra-class training scheme while VFS is an inter-class smoothing technique. This is a basic technique for on-line adaptation which will be important in constructing a practical speech recognition system. Speaker adaptation speed of the incremental MAP is experimentally shown to be significantly accelerated by the use of VFS in word-by-word adaptation. The recognition performance of MAP is consistently improved and stabilized by VFS. The word error reduction rate achieved in incrementally adapting a few words of sample data is about 22%.

1. INTRODUCTION

From recent intensive studies of Maximum a posteriori (MAP) estimation, or Bayesian Learning, for Hidden Markov Model (HMM), it has been reported that MAP estimation was experimentally proven to be effective for speaker adaptation [1, 2, 3, 5]. This is because the learning scheme of incorporating prior information into HMM training process is significantly useful for solving the inherent adaptation problem of insufficient training data. In addition, on-line adaptation which is not required in prior training data acquisition has been shown to have potential as an important and attractive application of the incorporated training function [4]. According to some of these studies, balance between the amount of adaptation training data and the number of models is suggested to be important in MAP adaptation. Thus, training speech lasting anywhere from ten seconds to several minutes, which correspond to the speech of one to several sentences, was often used. This volume of training data, however, is still thought to be too large to be collected in practical speech recognition applications. This results in a MAP adaptation speed that is too slow for practical applications.

This paper proposes a fast and incremental speaker adaptation method called MAP/VFS to improve the performance of MAP and to increase adaptation speed. This method combines MAP estimation as intra-class training with Vector Field Smoothing (VFS) [12] as inter-class smoothing. This concept of MAP/VFS first appeared in our work [6] in which this method was applied to simultaneous adaptation of a telephone line and a speaker and then the effectiveness was experimentally proven. In this study, MAP was used as a pooled training instead of Maximum Likelihood (ML) estimation. From the aspect of practical application, an incremental MAP/VFS was proposed in our further studies [7, 9, 11] for adaptation of telephone lines and of speakers to achieve fast and word-by-word adaptation. In these studies, how to build VFS into an incremental MAP procedure in order to expand MAP/VFS function toward on-line adaptation was investigated. After that, a speaker adaptation method based on the same presented combination was proposed in [8, 10]. In this paper, how to determine the prior density parameter in MAP/VFS method for fast and higher performance is described.

In the following section, the principle of MAP/VFS is described and the constituent algorithms of MAP estimation and VFS are also mathematically summarized. The remaining sections report some experimental results for word-by-word incremental speaker adaptation using the proposed method.

2. INCREMENTAL MAP/VFS ALGORITHM

2.1. MAP Estimation

The feature of MAP estimation is that the training incorporates prior information of model parameters and new incoming training data. This is expected to be an effective solution to the inherent problem of limited data training in adaptation. According to the mathematical estimation formulas for continuous-mixture-density HMMs, the mean vector of each output probability distribution, is estimated as follows.

$$\hat{\mathbf{m}}_k = \frac{\tau_k \mathbf{m}_k + \sum_{t=1}^{T} c_{kt}\mathbf{x}_t}{\tau_k + \sum_{t=1}^{T} c_{kt}}, \text{ where } c_{kt} = \frac{w_k N(\mathbf{x}_t \mid \hat{\mathbf{m}}_k, \boldsymbol{\Sigma}_k)}{\sum_k w_k N(\mathbf{x}_t \mid \hat{\mathbf{m}}_k, \boldsymbol{\Sigma}_k)} \tag{1}$$

Here, $\hat{\mathbf{m}}_k$ and $\mathbf{m}_k$ denote the MAP trained mean vector and the initial mean vector of the k-th mixture components, respectively. $\boldsymbol{\Sigma}_k$ represents covariance matrices. The parameter τ_k is a prior density parameter that controls the balance between prior information and the new training data. Equation (1) shows in such a way that the estimated mean vector is obtained as the interpolated one using the initial mean weighted by τ_k and the sample mean weighted by the effective sample counts $\sum_{t=1}^{T} c_{kt}$. As the ratio of τ_k against

0-7803-2431-5/95 $4.00 © 1995 IEEE

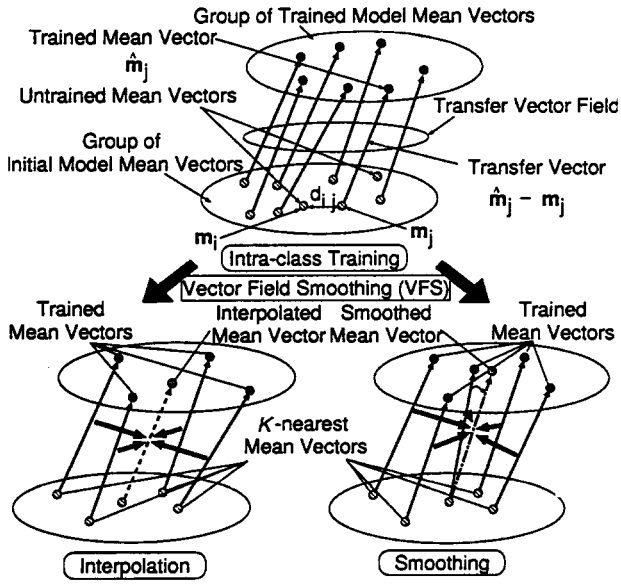

Figure 1. Illustration of Vector Filed Smoothing (VFS)

the sample counts approaches zero, the estimated mean approaches the sample mean from the initial mean. Accordingly, parameter τ_k is regarded as the control parameter for the adaptation speed. In incremental MAP, since the mean vector estimation is carried out one by one using training data, τ_k is renewed each time by $\tau_k \longleftarrow \tau_k + \sum_{t=1}^{T} c_{kt}$. With this parameter renewal of τ_k, the final estimated mean obtained from the incremental MAP procedure is theoretically equal to the estimated mean of the MAP procedure used as the pooled training.

2.2. Vector Field Smoothing (VFS)

VFS was proposed as one of solutions for insufficient training using a limited amount of sample data in speaker adaptation. The basic idea of VFS is that it is assumed that a set of training trajectories of mean vectors forms a smooth vector field.

The principle of VFS is illustrated in Figure 1. First, the initial mean vectors of initial models are trained using adaptation training data by an intra-class training method, such as Maximum Likelihood (ML) estimation or MAP estimation, in each of phone classes. Only partial mean vectors are trained because the amount of training data is often so small that they do not always include the data for training all initial models. Next, two kinds of estimations are carried out. One is interpolation to obtain the estimated mean vectors corresponding to the mean vectors left untrained in the initial models. The other is smoothing for the trained mean vector modification in order to reduce the estimation error caused by insufficient data intra-class training. In both cases, the mathematical formula is described in the following single equation.

$$\tilde{\mathbf{m}}_i = \frac{\sum_{j \in G_K} w_{ij}(\hat{\mathbf{m}}_j - \mathbf{m}_j)}{\sum_{j \in G_K} w_{ij}} + \mathbf{m}_i, \text{ where } w_{ij} = \exp(-d_{ij}^2/s) \tag{2}$$

In equation (2), $\mathbf{m}_j$ is the j-th initial mean vector and $\hat{\mathbf{m}}_j$ and $\tilde{\mathbf{m}}_j$ represent the j-th trained and adapted mean vector corresponding to $\mathbf{m}_j$. $(\hat{\mathbf{m}}_j - \mathbf{m}_j)$ is transfer vector in equation (2). The vector distance d_{ij} is the distance between i-th and j-th mean vectors, in which the Euclid distance measure is usually used. w_{ij} is the weighting coefficient for the neighboring transfer vectors included in G_K which denotes the group of K nearest-neighbor mean vectors. This value is given by the Gaussian window function shown in (2), where s is the smoothing parameter for adjusting the transfer vector direction. Since the i-th mean vector is not contained in G_K in interpolation, K nearest-neighbor mean vectors are used for estimating. In the smoothing procedure, the i-th mean vector is contained in G_K and then this vector and $K-1$ nearest-neighbor mean vectors are used for estimation error compensation. Parameter s is used to control the degree of smoothness in the transfer-vector field. If $s = 0$, for example, smoothing does not occur. Contrarily, if $s \to \infty$, all transfer vectors turn to the same direction when G_K contains all initial mean vectors.

2.3. Concept of MAP/VFS

MAP estimation is regarded as intra-class training in which the models can be trained only when the sample data to train them are acquired. On the other hand, VFS is regarded as inter-class smoothing. Therefore, the important concept is that the proposed MAP/VFS consists of two algorithms which compensate for each other's training function in order to yield a multiplying effect using each significant feature. In this study, another important point is made which is that the MAP estimation is adopted instead of Maximum likelihood (ML) estimation as intra-class training. (In the original VFS algorithm, ML was used). This is because, by the incorporated training which is equivalent to the training using a large amount of data, the MAP trained mean vectors are so well-trained that the well-estimated mean vectors will be obtained using VFS.

Let us consider the relation between MAP estimation and VFS based on the previously described mathematical formula. Equation (1) is rewritten as follows:

$$\hat{\mathbf{m}}_k - \mathbf{m}_k = \frac{\sum_{t=1}^{T} c_{kt}}{\tau_k + \sum_{t=1}^{T} c_{kt}}(\bar{\mathbf{x}}_t - \mathbf{m}_k), \text{ where } \bar{\mathbf{x}}_t = \frac{\sum_{t=1}^{T} c_{kt}\mathbf{x}_t}{\sum_{t=1}^{T} c_{kt}} \tag{3}$$

In equation (3), $\bar{\mathbf{x}}_t$ represents the sample mean of the adaptation training data. In the first equation in (3), the left term represents the transfer vector of MAP/VFS, while $(\bar{\mathbf{x}}_t - \mathbf{m}_k)$ in the right term corresponds to the transfer vector of ML/VFS because the sample mean $\bar{\mathbf{x}}_t$ is regarded as the ML estimated mean vector. Accordingly, in MAP/VFS, it is found that the transfer vector is weighted by the amount of training data. The weight is controlled by balancing between the sample counts and the amount of prior information represented by τ_k. In other words, the magnitude of the transfer vectors depends on τ_k. According to the incremental MAP principle, especially in the word-by-word speaker adaptation procedure, the MAP estimated mean vectors are thought to approach slowly the true mean

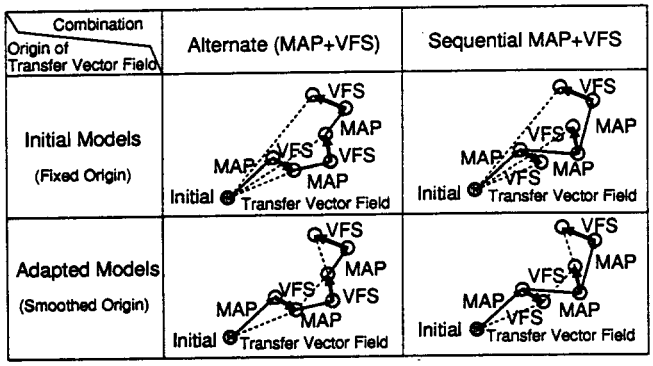

Figure 2. Four types of incremental MAP/VFS procedures

vectors as τ_k is iteratively renewed. To increase the adaptation speed in incremental MAP/VFS, it is effective to give τ_k a value rather less than the optimal value in the MAP adaptation.

3. ADAPTATION PROCEDURE

Figure 2 shows four possible combinations of MAP estimation and VFS. Each combination is different from each other in two ways. The first is in how VFS is incorporated into the incremental MAP procedure: *Alternate (MAP+VFS)* or *Sequential MAP+VFS*. In the former, the models left untrained are estimated by VFS using the MAP trained models, and the obtained models are then used as seed models for the next MAP estimation. This procedure is iteratively carried out each time that adaptation training data are acquired. Contrarily, in the latter, VFS is used as a post-processing of incremental MAP estimation. The MAP trained models are always used as seed models in the incremental procedure. The other way in which each combination differs is in which type of origin for the transfer-vector field is determined: *fixed origin* or *smoothed origin*. The origin is fixed to the initial models in the former while it is iteratively replaced by smoothed models using VFS in the latter. In the preliminary experiment, "*Alternate (MAP+VFS)* with a *fixed origin*" performed best and was used as the incremental MAP/VFS adaptation in this study.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Experiments

The recognition performance of the proposed MAP/VFS method was evaluated for speaker adaptation through phoneme recognition and a large vocabulary word recognition. These experiments were conducted using the speaker-independent recognition system that we developed. The initial models were context-dependent speaker-independent models [13]. These models are trained based on the Hidden Markov Network (HMnet) using a database, provided by ATR, which consists of 216 phonetically balanced word utterances and a half set of 5240 word utterances (even numbers) by 16 speakers (8 males and 8 females). This HMnet was constructed using allophone environment tying technique at triphone-model and state levels. There were 1498 context-dependent phoneme models in the tied-state configuration with two mixtures in output distributions. The number of states was 450 and the total output distribution counts were 924. The feature parameter was a 33-dimension vector consisting of 16 cepstral coefficients, 16 Δ cepstral coefficients, and a Δ log-power, which was obtained through LPC analysis under a sampling frequency of 12 [kHz], a Hamming window length of 32 [ms], and a frame shift of 8 [ms]. The adaptation training data and test data were selected from the other half set of the 5240 word utterances (odd numbers). 20 words were randomly selected as the adaptation training data and the remaining 2575 words were used as the test set. The test-set speakers were different from those of the training data set. In the experiments, the mean vector was adapted.

The recognition performance of MAP/VFS and incremental MAP were evaluated for two males and two females through phoneme recognition and a large vocabulary word recognition. A test set consisting of 2575 word utterances were used for both experiments. The vocabulary size was 2575 in isolated word recognition. Experiments were focused on investigating how the performance improvement and adaptation speed are changed by the value of the parameter τ_k in word-by-word incremental speaker adaptation. Two values of 16 and 7 for τ_k were used and the value of smoothing parameter s was fixed to 10.0.

4.2. Results

Figures 3 and 4 show adaptive learning curves of the proposed MAP/VFS and incremental MAP. The average recognition rates are plotted along the training word counts from 1 to 20. The performance of MAP/VFS and MAP in the case of $\tau_k = 7$ is compared in Table I for phoneme recognition and in Table II for a large vocabulary isolated word recognition. In the case of $\tau_k = 16$, the value of τ_k was iteratively renewed in both MAP/VFS and MAP adaptation according to incremental MAP algorithm. The performance of MAP/VFS increased rapidly in both phoneme and word recognition. For the incremental MAP, although the performance began to increase slowly from about four words in phoneme recognition, the word recognition performance was nearly equal to the baseline of the initial models. In another case where $\tau_k = 7$, τ_k was fixed to 7 through incremental adaptation in MAP/VFS and MAP. The performance of MAP/VFS significantly more improved over that of $\tau_k = 16$ in phoneme and word recognition. Nevertheless, in incremental MAP, the phoneme recognition performance decreased and the word recognition performance tended to degrade slowly to beneath the baseline. From these results, a τ_k-value less than the optimal value of MAP was experimentally proven to be good for MAP/VFS. This is because, in the case of $\tau_k = 7$, adaptation speed is thought to be accelerated by means in which a proper over-estimation of mean vectors in MAP is carried out to obtain as long a transfer vector as possible since the transfer vector is short in word-by-word adaptation and the variance of the over-estimated mean vectors is adjusted by the smoothing of VFS. This suggests that the optimal value of τ_k may be different between MAP/VFS and MAP. With the proposed MAP/VFS, the phoneme error was reduced by 7.4%, 11.5%, and 14.7% for 4-word, 10-word, and 20-word adaptation, respectively. The word error was also comparatively reduced by 16.5%, 21.6%, and 22.7%. The results show that not only did VFS consistently accelerate the adaptation speed

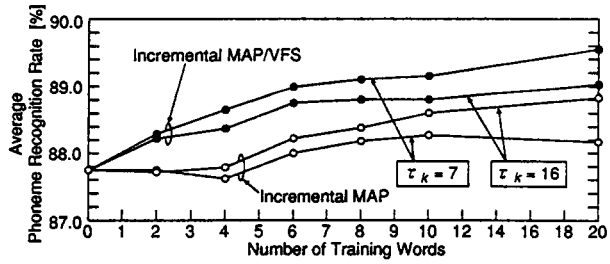

Figure 3. Adaptive learning curves in phoneme recognition

Table I. Phoneme recognition performance comparison

Speaker	Adaptation Method	Number of Training Words: 0	4		10	
		Recognition Rate [%]	Recognition Rate [%]	*Error Reduction Rate [%]*	Recognition Rate [%]	*Error Reduction Rate [%]*
MMY	MAP/VFS	83.9	86.4	*15.5*	87.7	*23.6*
	MAP		84.1	*1.2*	86.1	*13.7*
MMS	MAP/VFS	87.9	88.4	*4.1*	87.6	*-2.5*
	MAP		87.9	*0.0*	87.1	*-6.6*
FYN	MAP/VFS	89.7	90.5	*7.8*	91.1	*13.6*
	MAP		89.6	*1.0*	90.4	*6.8*
FKS	MAP/VFS	89.5	89.3	*-1.9*	90.2	*6.7*
	MAP		88.9	*-5.7*	89.5	*0.0*
Average	MAP/VFS	87.8	88.7	*7.4*	89.2	*11.5*
	MAP		87.6	*-1.6*	88.3	*4.1*

of the incremental MAP but also improve and stabilize performance by combining MAP with VFS.

5. CONCLUSION

In this paper, we presented the concept of MAP/VFS adaptation method to achieve fast and incremental speaker adaptation. In the performance evaluation, the adaptation speed of the incremental MAP-based speaker adaptation is shown to be consistently accelerated by the Vector Field Smoothing (VFS) even if the amount of training data was very small. Furthermore, VFS improves and stabilizes the recognition performance of the incremental MAP. The proposed method is simple to process and requires no pooling of the adaptation training data.

ACKNOWLEDGEMENT

The authors would like to thank Drs. Nobuhiko Kitawaki and Noboru Sugamura for their administrative supports in this work. We would also like to acknowledge Satoshi Takahashi for his helpful technical supports.

REFERENCES

[1] 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. ASSP*, Vol. 39, No. 4, pp. 806-814, April 1991.

[2] J. -L. Gauvain and C. -H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc. DARPA Speech and Natural Language Workshop*, pp. 272-277, Arden House, Feb. 1991.

[3] C. -H. Lee and J. -L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," *Proc. ICASSP93*, pp. II-558-561, Apr. 1993.

[4] T. Matsuoka and C. -H. Lee, "A Study of On-line Bayesian Adaptation for HMM-based Speech Recognition," *Proc. EUROSPEECH93*, pp. 815-818, 1993.

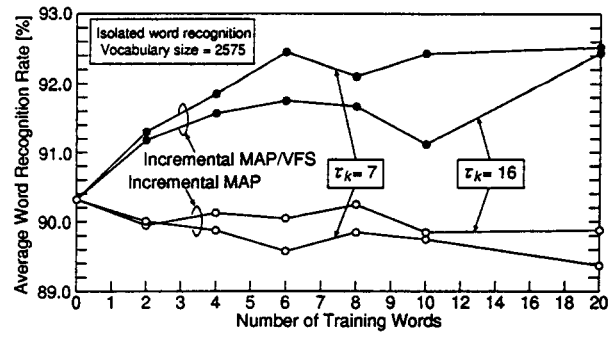

Figure 4. Adaptive learning curves in word recognition

Table II. Word recognition performance comparison

Speaker	Adaptation Method	Number of Training Words: 0	4		10	
		Recognition Rate [%]	Recognition Rate [%]	*Error Reduction Rate [%]*	Recognition Rate [%]	*Error Reduction Rate [%]*
MMY	MAP/VFS	85.8	87.7	*13.4*	90.8	*35.2*
	MAP		86.2	*2.8*	88.5	*19.0*
MMS	MAP/VFS	91.2	92.7	*17.0*	92.0	*9.1*
	MAP		91.6	*4.5*	90.0	*-13.6*
FYN	MAP/VFS	93.1	94.3	*17.4*	94.6	*21.7*
	MAP		92.0	*-15.9*	90.8	*-33.3*
FKS	MAP/VFS	91.2	92.7	*17.0*	92.3	*12.5*
	MAP		89.7	*-17.0*	89.7	*-17.0*
Average	MAP/VFS	90.3	91.9	*16.5*	92.4	*21.6*
	MAP		89.9	*-4.1*	89.8	*-5.2*

[5] Y. Turumi and S. Nakagawa, "An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum A Posteriori Probability Estimation," *Proc. ICSLP94*, pp. 431-434, Sept. 1994.

[6] J. Takahashi and S. Sagayama, "Telephone Line Characteristic Adaptation Using Vector Field Smoothing Technique," *Proc. ICSLP94*, pp. 991-994, Sept. 1994.

[7] J. Takahashi and S. Sagayama, "Fast Telephone Channel Adaptation Based on Vector Field Smoothing Technique," *Proc. IVTTA94*, pp. 97-100, Sept. 1994.

[8] M. Tonomura, T. Kosaka and S. Matsunaga, "Unified Speaker Adaptation Algorithm of Maximum a Posteriori Estimation and Transfer Vector Field Smoothing," *Technical Report of IEICE*, SP94-51, pp. 25-30, Oct. 1994 (in Japanese).

[9] J. Takahashi and S. Sagayama, "Vector-Field-Smoothed Bayesian Learning for Fast Speaker Adaptation," *Proc. Fall Meeting, Acoustical Society of Japan*, 2-8-19, pp. 75-76, Nov. 1994 (in Japanese).

[10] M. Tonomura, T. Kosaka and S. Matsunaga, "Speaker Adaptation Based on Transfer Vector Field Smoothing Model Using Maximum A Posteriori Probability Estimation," *Proc. Fall Meeting, Acoustical Society of Japan*, 2-8-20, pp. 77-78, Nov. 1994 (in Japanese).

[11] J. Takahashi and S. Sagayama, "Vector-Field-Smoothed Bayesian Learning for Incremental Speaker/Telephone-Line Adaptation," *Technical Report of IEICE*, SP94-74, pp. 33-40, Dec. 1994 (in Japanese).

[12] K. Ohkura, M. Sugiyama, and S. Sagayama, "Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," *Proc. ICSLP92*, pp. 369-372, Oct. 1992.

[13] S. Takahashi and S. Sagayama, "Four-Level Tied-Structure for Efficient Representation of Acoustic Modeling," to appear in *Proc. ICASSP95*, Detroit, May 1995.