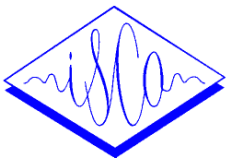

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

A DYNAMIC APPROACH TO SPEAKER ADAPTATION OF HIDDEN MARKOV NETWORKS FOR SPEECH RECOGNITION

Tetsuo Kosaka[1], Edward Willems[2], Jun-Ichi Takami[1] and Shigeki Sagayama[3]

[1]*ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-02 Japan*
[2]*Ecole Nationale Superieure des Telecommunications, 46 rue Barrault 75634 Paris Cedex 13 France*
[3]*NTT Human Interface Research Labs., 3-9-11 Midori-cho Musashino-shi Tokyo 180 Japan*

ABSTRACT

This paper describes a new approach to dynamic speaker adaptation, which relies on switching between different methods of adaptation in order to gain maximum performance depending on the amount of speech data obtained through the speech recognition session. This adaptaion method has been successfully applied to a hidden Markov network (HMnet), which is an efficient representation of phoneme context-dependent HMMs. This speaker adaptation method has proven itself effective in improving the performance of a Japanese phrase recognition system.

Keywords: *speech recognition, speaker adaptation, hidden Markov network*

1. INTRODUCTION

Various techniques of speaker adaptations have been developed to reduce the differences in performance between speaker-dependent systems and speaker-independent systems [1][2][5][6][7]. For practical applications, it is desirable that a speaker adaptation method requires only a small amount of training data.

In order to reduce a speaker's adaptation cost, we plan to develop a rapid on-line adaptation system(figure 1).

The system consists of a model-adaptation module and a performance-evaluation module. The first uses any available technique to adapt a model to the current input speaker. The quality of this model is evaluated using the available training data. The results of this evaluation are used to control the adaptation module. There is therefore a feedback loop between the modules. The evaluation module also gives a recognition result. Accordingly the model is dynamically modified by current input data in this system; we call this *a dynamic speaker adaptaion method.* In such a dynamic adaptation system, it is desirable to achieve the highest possible increase in recognition with as little training data as possible.

Speaker adaptation method based on maximum a posteriori(MAP) estimation[3] has been proposed recently to enable such a dynamic adaptation. It is possible to incorporate this MAP-based adaptation in our novel adaptation method which is described below.

In general, adaptation methods involve the following two trade-offs:

- An adaptation method with a small number of free parameters for adaptation requires less training data, but its final recognition rate is lower.
- An adaptation method with a large number of free parameters for adaptation requires more training data, but its final recognition rate is higher.

With these trade-offs in mind, we propose a novel adaptation method by combining different adaptation methods having different numbers of free parameters to achieve a rapid speaker adaptation in the dynamic adaptation system.

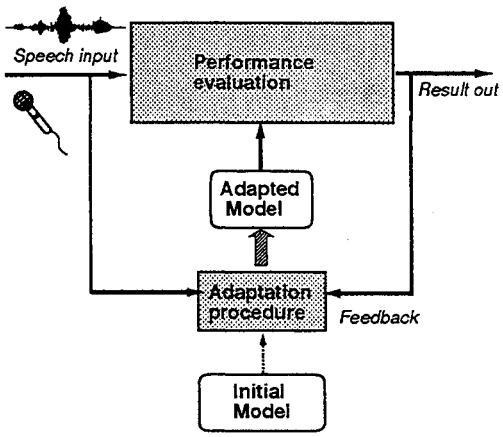

Figure 1. Block diagram of a dynamic speaker-adaptive system

2. PRINCIPLE

The present system can switch between three different speaker adaptation methods: vector field smoothing (VFS)[5], speaker-tied weight training (STWT)[6] and speaker-free weight training (SFWT)[7]. These methods are described below.

These methods have a different number of free parameters. These quantities are as follows:

$$STWT < SFST < VFS$$

Consequently, we can predict that in general the final recognition rate will be lowest for STWT and highest for VFS. We can also predict that the former method requires less data than the latter. It therefore follows that selection of adaptation methods according to the amount increase in training data realizes a rapid speaker adaptation.

The efficiency of any adaptation technique is highly dependent on the speaker's phonetic characteristics. Thus it is impossible to determine in advance the regions within

which a given method is most effective. To solve this problem, we propose an automatic selection of the most effective method.

The methods are selected according to the most likely candidate they would produce, based on the input speech data.

3. ALGORITHM

The observation is a sequence of n-dimensional vectors $\mathbf{X} = \{x_1, \ldots, x_K\}$. $m_i k$ is a phone model set (e.g. hidden Markov network[4] described below), which is trained by $\{x_1, \ldots, x_k\}(1 \leq k \leq K)$ using adaptation method i.

Dynamic speaker adaptation is carried out in the following three steps:

Speaker adaptation: An initial phone model set is adapted by input $\{x_1, \ldots, x_{k-1}\}$ using several speaker adaptation methods. After adaptation, phone model sets $\{m_{1k-1}, \ldots, m_{Ik-1}\}$ are obtained, where I is the number of adaptation methods. We used a speaker-independent phone model set trained by the speaker-mixture argorithm as an initial model set.

Automatic model selection: One of the adapted model sets is selected automatically based on a maximum likelihood criterion.

$$M_{k-1} = \underset{i}{\operatorname{argmax}}\, L(x_k, m_{ik-1}) \qquad (1)$$

where M_{k-1} is the selected model set, and $L(x_k, m_{ik-1})$ is an output likelihood derived from m_{ik-1} for input x_k.

Recognition: Input x_k is recognized by using selected model set M_{k-1}.

4. SPEAKER MIXTURE

We have already proposed a speaker-mixture method[6] that can yield highly accurate speaker-independent phone models. We use these speaker-independent phone models in this paper.

Suppose that a set of speaker-dependent continuous mixture HMM phone models is given for each of several reference speakers. Speaker-mixture phone models are constructed by merging all corresponding states of each speaker with equal speaker weights. This is a kind of hierarchical mixture model that contains speaker mixture weights and intra-speaker mixture weights for mixture components.

In continuous mixture HMMs, the probability density function $b_j(\mathbf{x})$ associated to state j, $1 \leq j \leq N$, can be written as:

$$b_j(\mathbf{x}) = \sum_{s=1}^{S} w_j^{(s)} b_j^{(s)}(\mathbf{x}) \qquad (2)$$

where x is the input vector, s is the mixture number, S is the total number of mixtures in one state, and $w_j^{(s)}$ is the mixture weight. In the speaker mixture method, each mixture component $b_j^{(s)}$ is derived from speaker s for all states j. Thus S is set to the number of speakers.

Rapid speaker adaptation using speaker-tied mixture-weight training is derived from this principle.

The principle of speaker mixture is illustrated in figure 2 (a).

5. ADAPTATION METHODS FOR DYNAMIC SPEAKER ADAPTATION

The dynamic speaker adaptation described in this paper requires several adaptation methods that have different numbers of free parameters. Three adaptation methods (VFS[5], STWT[6] and SFWT[7]) are used for this dynamic speaker adaptation.

5.1. Speaker-tied mixture weight training(STWT)

The available training data are used to determine which mixture-components are closest to the input speaker phones in the speaker-mixture model. This is done using the Baum-Welch algorithm. The initial equal weights are then modulated so as to favor the ones close to the input phones. The adapted HM-Net is then the weighted combination of the mixture components, based on the training data values. The principle of speaker-tied mixture weight training is illustrated in figure 2 (b).

5.2. Speaker-free mixture weight training(SFWT)

This technique relies on weighting the mixture components to adapt the HM-Net. However, whereas in the previous case a weight was given to each component and applied to all states of the HM-Net, the weight parameters are here calculated separately for each state. The weights are determined by the Baum-Welch algorithm. The states of the adapted network are therefore made up of the weighted combination of the equivalent states in the mixture components. However, the weighting is calculated according to the training data values, and is therefore only defined for the states in the network which were accessed by the training data. The adaptation is therefore only carried out for these states, and the others are left as an unbiased combination of the mixtures. The principle of speaker-free mixture weight training is illustrated in figure 2 (c).

5.3. Vector Field Smoothing adaptation(VFS)

In the VFS algorithm, the mapping from initial HMMs to retrained HMMs, i.e. from the reference speaker to a new speaker, is carried out according to a transfer vector field. Two steps are carried out in the VFS algorithm:

Embedded Training: The mean vectors of the Gaussian distribution in the HMnet are trained by embedded training using the training data and its phonetical transcription.

Smoothing of Transfer Vectors: The transfer vectors used for the adapted mean vectors are smoothed by a spatial filtering technique. This is to estimate mean vectors that have not been adapted because their corresponding phoneme context does not appear in the training samples. This step also corrects the estimation errors of mean vectors resulting from a small amount of training samples.

The smoothing rate is a variable in order to control the strength of the smoothing. This rate corresponds

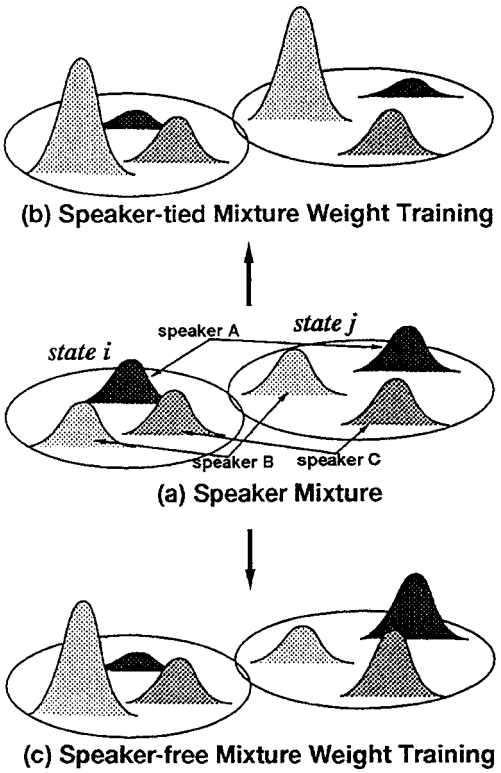

Figure 2. Principle of speaker-mixture

to the window width of the spatial filter. Stronger smoothing is possible by increasing the smoothing rate.

6. HIDDEN MARKOV NETWORKS

A dyanamic speaker adaptaion method has been successfully applied to a hidden Markov network (HMnet)[4], which is an efficient representation of phoneme context-dependent HMMs.

The HMnet is a highly generalized form of the HMM, and incorporates context-dependent variations of phones and state sharing among different allophones. The HMnet contains a finite number of states, each containing Gaussian distributions, that are connected to each other to form paths representing context-dependent phone models. This network is automatically derived by using the Successive State Splitting (SSS) algorithm, which simultaneously solves three problems: network topology, allophone clusters, and the acoustic distribution for each state. Figure 3 shows the structure of the HMnet.

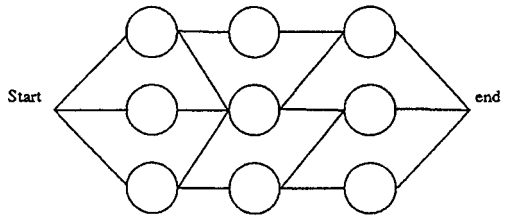

Figure 3. Structure of the Hidden Markov Network

7. RECOGNITION EXPERIMENTS

This dynamic adaptation method was tested on Japanese phrase recognition. The experimental conditions are summarized in table 1, and the evaluation system is schematically shown in figure 4.

The 12-mixture 200-state HMnet was trained with 216 isolated words uttered by 12 male speakers. A speaker-mixture SSS algorithm is used for training.

These models were combined with a generalized LR parser [8], which could cope with a context-free grammar, to recognize Japanese phrases. The task included 1,000 words and its phoneme perplexity was 5.9.

Experiments have been conducted on two dialogues concerning "inquiry about an international conference registration." Input sentences were spoken by two spekers and uttered phrase by phrase. The number of phrases in each dialogue is 256 (SB1 data set) and 279 (SB3 data set). The adaptation data was the SB1 phrase set and the testing data was the SB3 phrase set.

Table 1. Experimental Conditions

Analysis Conditions		
Sampling-rate	12kHz	
Window	Hamming window (20ms)	
Frame period	5ms	
Analysis	log power + 16-order LPC-Cep + Δlog power + 16-order ΔLPC-Cep	
Training data		
for structure generation	1 male	5240 words
for parameter training	12 males	216 words
Adaptation and recognition data		
Speaker	2 males	
Adaptation data	256 phrases (SB1 task)	
Recognition data	279 phrases (SB3 task)	

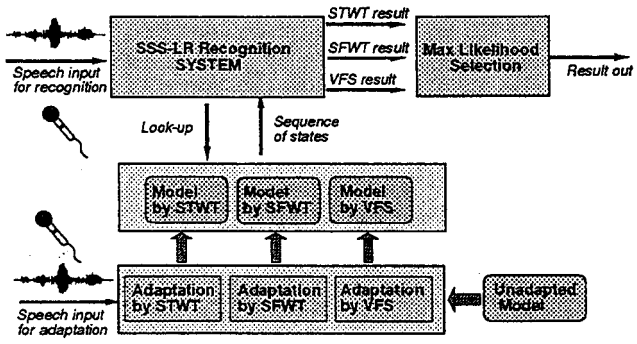

Figure 4. Block diagram of an evaluation system

8. DISCUSSION

Figure 5 indicates recognition rates achieved when using the four adaptation techniques with various amounts of training data. The results show that the efficacy of each speaker adaptation method depends on the speakers.

For speaker MSH, STWT was the most effective method while using a small amount of training data; however, the highest performance could be obtained with a larger

amount of training data using VFS. As evident in this figure, the overall performance of the dynamic adaptation is better than that of any other single technique. The recognition rate achieved is higher than that of VFS by using 50 utterances. This is simply explained by the fact that some of the errors of VFS are correctly deciphered by one of the other two techniques.

For speaker MTM, VFS was the most effective method, while STWT was not effective at all. These results suggest that among the reference speakers, who were used to train the original speaker mixture model, no one was acoustically "close" to the input speaker.

The reason why the dynamic method was not most effective for speaker MTM was that the re-trained parameter types were not the same for VFS and weight training adaptation methods. For VFS, output probability density functions are used in the retraining of the mean vectors, whereas, for STWT and SFWT, probabilities are used in the retraining of the mixture weights. The dynamic range for probability density vs. probability differs considerably and cannot be compared directly.

The precise rates of method selection is represented in figure 6. The diagram shows that as the number of training samples is increased, the STWT method is less often selected, while the selection rates for the other two methods increases.

In comparing the two speakers, VFS was more frequently selected for speaker MTM, because it is more effective for that speaker.

9. CONCLUSIONS

A new approach for rapid speaker adaptation using the technique of combining different adaptation methods has been proposed in this paper. In this approach, the system can automatically switch between several different adaptation methods by using the maximum likelihood criterion. The new method has been successfully tested on Japanese phrase recognition tasks. Future work will involve application of this method to unsupervised adaptation.

ACKNOWLEDGMENTS

We would like to thank Dr. Yamazaki, President, ATR Interpreting Telecommunications Research Laboratories, and Dr. Kurematsu, Professor, University of the Electro-communications, for their continuous support of this work. We are also grateful to all of the members of Department 1 for their advice and encouragement.

REFERENCES

[1] F. Kubala, et al.: "Speaker Adaptation from a Speaker-Independent Training Corpus," Proc. of ICASSP'90, pp. 137-140 (1990).
[2] X. Huang and K. Lee: "On speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition," Proc. of ICASSP'91, pp. 877-880 (1991).
[3] C. -H. Lee and J. -L. Gauvain: "Speaker Adaptation Based on MAP Estimation of HMM Parameters," Proc. of ICASSP'93, pp. 558-561 (1993).
[4] J. Takami, et al.: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP'92, pp. 573-576 (1992).
[5] K. Ohkura, et al.: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of ICSLP'92, We.fPM.1.1, pp. 369-372 (1992).
[6] T. Kosaka, J. Takami and S. Sagayama: "Rapid Speaker Adaptation Using Speaker-Mixtue Allophone Models Applied to Speaker-Independent Speech Recognition," Proc. of ICASSP93, pp. 570-573 (1993)
[7] T. Matsuoka, K. Shikano: "Speaker Adaptation by Modifying Mixture Coefficients of Speaker-Independent Mixture Gaussian HMMs," Proc. of ICSLP'92, We.fPM.1.2, pp. 373-376 (1992).
[8] A. Nagai, et al.:"Phoneme-context-dependent LR Parsing Algorithms for HMM-based Continuous Speech Recognition," Proc. of Eurospeech'91, S48.3. (1991).

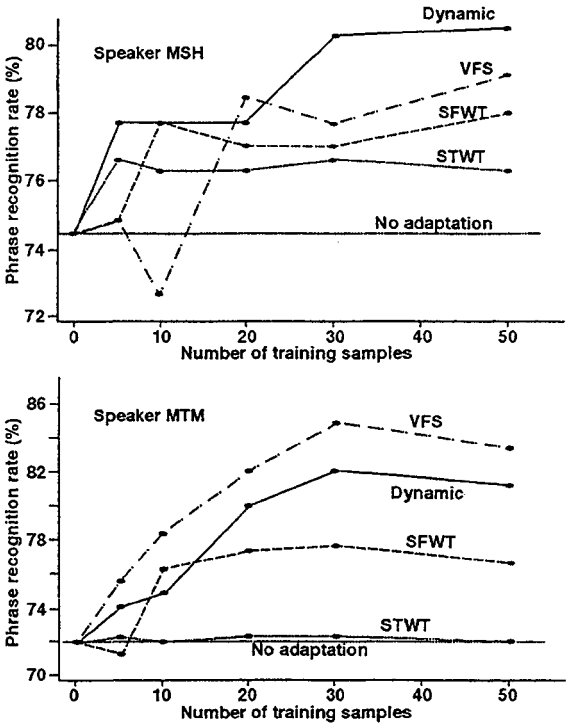

Figure 5. Performance evaluation of speaker adaptation methods

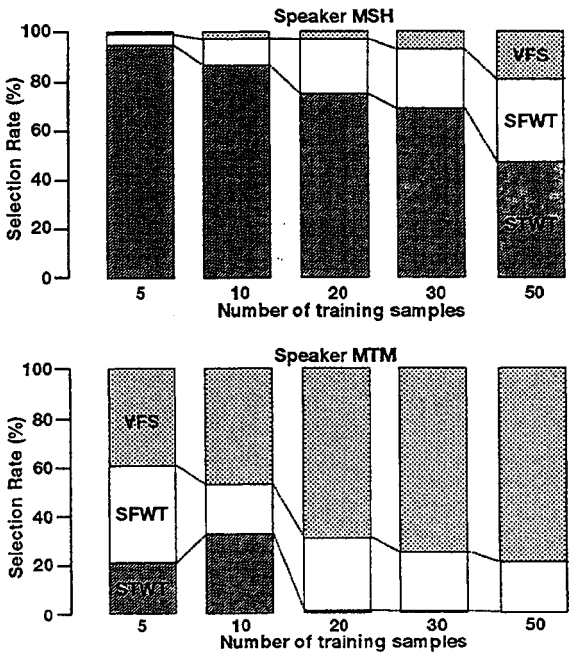

Figure 6. Precise rates of method selection