論文 / 著書情報
Article / Book Information

Title	Speaker adaptation with autonomous control using tree structure
Authors	K. Shinoda, T. Watanabe
Citation	Proc. EuroSpeech-95, Vol. , No. , pp. 1143-1146
Pub. date	1995,

SPEAKER ADAPTATION WITH AUTONOMOUS CONTROL USING TREE STRUCTURE

Koichi Shinoda and Takao Watanabe

Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
e-mail: shinoda@hum.cl.nec.co.jp

ABSTRACT

In practical use of speaker adaptation, it is important to provide a framework that operates well for any amount of adaptation data since the amount of data available is often changed. We propose one such framework in which the number of free parameters for estimation is autonomously controlled according to the amount of data for adaptation. It has been applied to a speaker-independent speech recognition system using continuous density mixture Gaussian HMMs, and has proven to be effective through 5,000-word recognition experiments. For example, it achieved a 40% reduction in error rate over the speaker-independent recognition system when 50 words were used for adaptation.

1. INTRODUCTION

Over the last few years, extensive studies have been carried out on Speaker-Independent(SI) speech recognition systems using continuous density mixture Gaussian Hidden Markov Models(HMMs). As the model parameters are estimated with a large amount of training data from many speakers, they are robust against speaker variety and show good performances on average. They have an advantage over Speaker-Dependent(SD) recognition systems in that they do not need any training data from a new speaker. However, their recognition accuracies are less than those obtained with well-trained SD recognition systems. Furthermore, they have shown poor performance for some particular speakers, probably because the acoustical properties of those speakers were different from those of any of the speakers from whom training data had been obtained. To overcome these problems, various speaker adaptation methods have been developed and applied to SI recognition systems.

In general, speaker adaptation methods are expected to have the following two properties.

- The recognition accuracy is improved with even a small amount of data for adaptation.
- The recognition accuracy increases as the amount of data increases, even when the amount of data available is large.

Many adaptation techniques have been developed, but few of them have fulfilled both requirements.

As means of achieving the former requirement, parameter-tying and parameter-smoothing techniques[1, 2, 3] have been studied. By reducing the degree of freedom with which free parameters can be estimated, these techniques improve recognition accuracies even when the amount of adaptation data is extremely small. However, the recognition accuracies stop increasing with a relatively small amount of adaptation data, and remains less than those of SD recognition when a large amount of data is available for adaptation. This is because the models adapted with the predetermined low degree of freedom fail to capture the rich acoustical characteristics that the large data contains. Thus, these techniques do not meet the latter requirement. It can be expected that these two requirements will be both fulfilled with techniques that gradually untie the parameters, or gradually reduce the smoothing effects with the increase in data. In this paper, we propose one such technique, in which the degree of freedom with which free parameters can be estimated is *autonomously* controlled according to the amount of data for adaptation. Utilizing a tree structure of parameters for this autonomous control, it provides a free parameter set of adequate size for any amount of adaptation data.

2. AUTONOMOUS CONTROL OF PARAMETER SET SIZE

Each state in continuous density mixture Gaussian HMMs has an output probability density function, which is the weighted summation of K Gaussian components. In the adaptation procedure, SI mean μ_{i0} of Gaussian component i is mapped to the unknown SD mean $\hat{\mu}_i$ as follows:

$$\hat{\mu}_i = \mu_{i0} + \delta_i, \qquad i = 1, \ldots, N \times K,$$

where δ_i is a shift parameter from the SI mean, and N is the number of states in HMMs. We attempt to estimate these shifts for all the Gaussian components in HMMs.

The total number of the Gaussian components in HMMs, $N \times K$, is so large that estimating the shifts with a small amount of data causes serious degradations in recognition accuracies. One way to solve this problem is to reduce the number of free parameters by tying the shifts. However, the models with the small number of free parameters do not give much improvement in recognition accuracies when a large amount of adaptation data is available. Therefore, the number of free parameters should be controlled according to the amount of adaptation data.

As means of controlling the number of free parameters, we introduce a tree structure for the shifts(Figure 1). In this tree structure, each leaf node i corresponds to each Gaussian component i, and a tied-shift Δ_j is defined for each non-leaf node j (including a top node). The shift δ_i for each leaf node i is estimated using a training algorithm such as Viterbi algorithm or Forward-Backward

algorithm. The tied-shift Δ_j for each non-leaf node j is calculated as a shift shared by all the leaf nodes that fall below the node j.

Using this tree structure, the number of free parameters can be controlled according to the amount of data. When the amount of data is small, the tied-shift Δ_j of the nodes at the higher level is applied to the Gaussian components of the leaf nodes that fall below the node j; it is used as the shifts for those leaf nodes. When the amount of data is large, Δ_j of the node at the lower level is used as the shifts δ_i. To autonomously control the number of parameters, we set a threshold on the amount of *node data*. Here *node data* is defined for each node j as the data used for estimating the tied shift Δ_j. Only the nodes for which there are moderate (neither too small nor too large) amount of node data are selected and their tied-shifts Δ_j are used for adaptation. This principle is depicted in Figure 1.

When the total amount of data for adaptation is very small, only the top node is selected for adaptation. The parameter Δ_j of the top node is applied to all the means in HMMs. This tie-shift represents a movement of the whole of the means in the acoustical feature space from the SI means. As the total amount of data increases, the nodes at the lower levels are selected for adaptation. The tied-shifts Δ_j of these nodes represent more local movements. When a sufficient amount of data is available, the shift δ_i of each leaf node i is directly used for the adaptation of the corresponding Gaussian component i. In this case, the adapted models become identical with the models adapted without the tree structure.

3. IMPLEMENTATION

The performance of the proposed method largely depends on how the tree structure is constructed. The tree should be made so that the tied-shift Δ_j of the node j well represents all shifts for the leaf nodes that fall below the node j. According to a recent research report[4], one can assume that, for the Gaussian components that are close to each other in the acoustical feature space, the shifts also tend to be close to each other. Therefore, we construct the tree using the distance between the Gaussian components in SI models. The tree is made in a top-down manner with the k-means algorithm, in which the divergence is used as the distance between the Gaussian components[5].

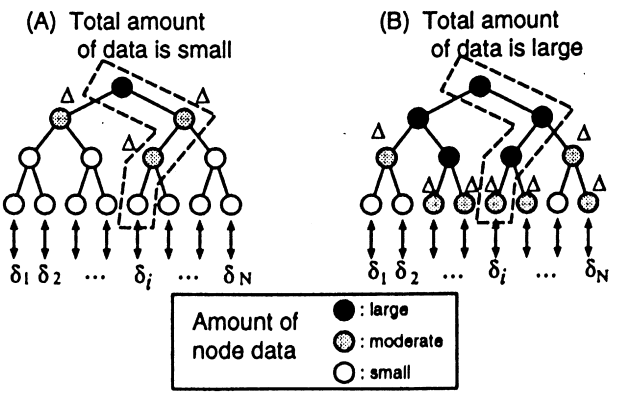

Figure 1: Principle

The proposed method is implemented into a supervised adaptation scheme. The adaptation procedure has the following four steps.

1. Each data frame in the adaptation data is assigned to a Gaussian component of the SI HMMs using the Viterbi time-alignment. The difference between the feature at the data frame and the mean is calculated, and is attached to the leaf node which corresponds to the Gaussian component. After processing for all the data, each leaf node i has a set of the differences $\{V(i,1),\ldots,V(i,K_i)\}$, where K_i is the number of data frames for node i.
2. The shift δ_i for each leaf node i is calculated as the average of $V(i,K_i); k = 1,\ldots,K_i$. The tied-shift Δ_j for the non-leaf node j is calculated as the average of the differences for all the leaf nodes that fall below the node j.
3. Let D be a threshold on the number of data frames. For each leaf node i, one node is selected: the selected node is either the leaf node i itself or one of the dominating nodes, which are surrounded by a dashed line in Figure 1. If the leaf node has more data frames than D, it is selected. Otherwise, the nearest node to the leaf node which has more data frames K_i than N is selected.
4. For each leaf node, the tied-shift of the selected node is added to the mean.

Let us compare the proposed method with some of the other techniques. Digalakis et al. recently proposed "genones" approach [2], in which the mixture components are tied to each other. Our method autonomously changes the degree of tying according to the amount of data available. We also compare our method with the Vector Field Smoothing (VFS) approach[3], in which the shifts are smoothed in acoustic feature space. In the VFS approach, the degree of smoothing is controlled by a function of the distances between Gaussian components. In our approach, the spatial relations among Gaussian components are represented by the tree structure.

4. EXPERIMENTS

4.1. Experimental Conditions

The proposed method was evaluated using the demi-syllable based Japanese recognition system[6]. In this system, speech was digitized at a 16 kHz sampling rate, and analyzed by a 10 msec frame period. The feature used was a vector of 21 components, consisting of a power derivative, 10 mel-scaled cepstral coefficients, and 10 corresponding mel-scaled cepstral time derivatives. The number of demi-syllable HMMs was 260, each of which had one to four states. The number of mixtures in each state was fixed at two. The diagonal covariance was used for each mixture component. The total number of Gaussian components was 2046.

The baseline SI models were trained on an database consisting of data from 85 speakers, in which each speaker uttered 250 phonetically-balanced words. The tree structure for the shifts was constructed using these SI models. The tree had five levels from top to bottom, numbered $L1$ through $L5$ respectively. From the top level $L1$ to the fourth level $L4$, there were four branches for each node. In the bottom level $L5$, each node had a different number

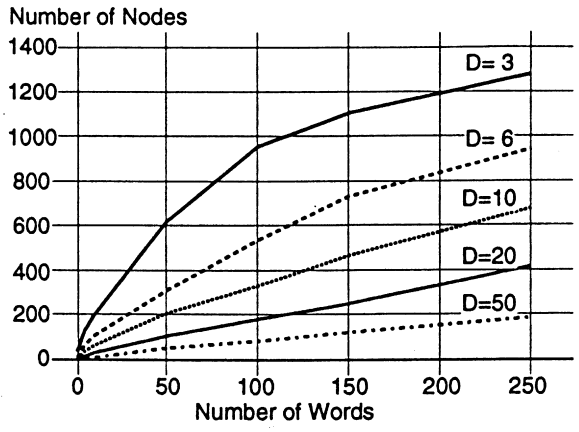

Figure 2: Number of nodes used for adaptation

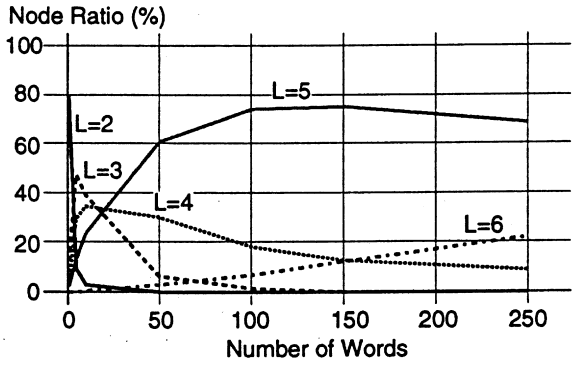

Figure 3: Ratio of number of nodes in each level of tree structure

of leaf nodes. The level for the leaf nodes was defined and was named $L6$.

The proposed method was evaluated with large-vocabulary, isolated-word recognition experiments. As a recognition vocabulary, a set of 5,000 words selected from a Japanese dictionary was used. For testing, the speech data of five male speakers and five female speakers, who were not involved in the training database, were used. Each of these test speakers uttered 250 words for adaptation and 250 words for testing. The vocabulary for testing was different from that for training and adaptation.

4.2. Results

Prior to the recognition experiments, we examined how the number of parameters was controlled using the proposed method. Figure 2 shows the number of nodes used for adaptation of male speaker M4, where D means the threshold for the number of data frames for each node. The number of nodes became larger in proportion to the number of words. Figure 3 graphs the ratio of the number of nodes used for adaptation in the tree's L levels to the total number of nodes used for adaptation, as a function of the number of words. The threshold D was fixed at 10. From the figure, one can see that as the number of words increases, the dominant level (the level achieving the highest ratio) approaches the leaf node level $L6$.

The recognition results of the proposed method, averaged for 10 speakers, are shown in Figure 4 in which the threshold D is varied from 3 to 50. In this figure, the SI recognition result(SI) and the result of the reference experiments(ref), in which the tree structure was not used for adaptation, are also shown. In the reference experiments, the spectral interpolation technique[4] and the MAP estimation of the means[7] were employed. As shown in the figure, the optimal value for the threshold was proved to be 10. With this threshold, the proposed method achieved a 40% reduction in error rate over the speaker-independent recognition system when 50 words were used for adaptation. It also achieved a better result than the method without the tree for any amount of data.

Figure 5 shows the recognition rates for the 10 speakers. It shows the results of SI experiments(SI), 10-word adaptation, and 250-word adaptation. The most significant improvement was seen for speakers with lower accuracy in the SI experiments. The rate of speaker M3 was not improved with 10-word adaptation. However, this was not a serious problem because the SI recognition rate for this speaker was high enough.

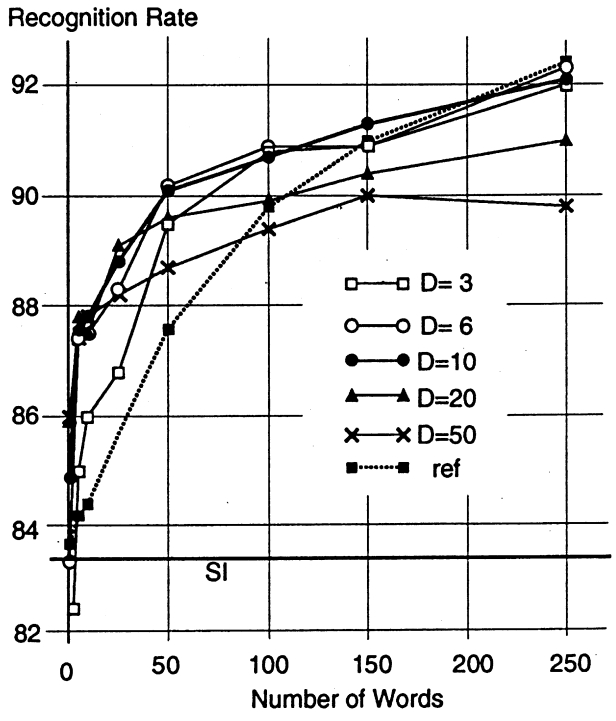

Figure 4: Proposed method

In Figure 6, the result of the experiments in which the level L in the tree used for adaptation was fixed(level-fixed method). In this figure, the result of the proposed method with the threshold 10 is also shown. When the number of words was small, the rate for the higher level was better than that for the lower level. As the number of words increased, the level which achieved the highest rate approached the leaf node level L6. The proposed method achieved higher rates than the level-fixed methods for any levels, regardless of the number of words. This means that the autonomous control of the parameter number

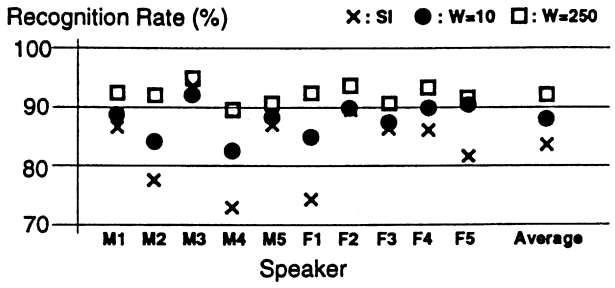

Figure 5: Recognition accuracy for each speaker

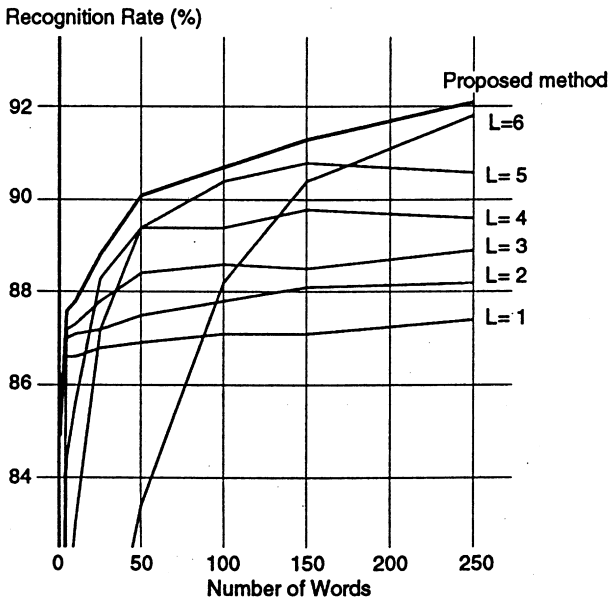

Figure 6: Recognition accuracy for each level

was successful.

As one can see from Figure 4, the optimal threshold value was different for each number of words. When the number of words was small, a high threshold value achieved high rates, but with a large number of words, a low threshold was preferable. This issue will be further investigated in future research.

5. CONCLUSION

A speaker adaptation method, in which the size of the parameter-set to be estimated is autonomously controlled using a tree structure for the parameters, has been proposed. This method has been proven to be effective in large vocabulary experiments. Future work will apply this method to the unsupervised, on-line adaptation scheme.

ACKNOWLEDGMENTS

The authors wish to thank the members of the Human Language Research Laboratory for their continuous support and encouragement.

REFERENCES

[1] J.Bellegarda and D. Nahamoo, " Tied Mixture Continuous Parameter Modeling for Speech Recognition", *IEEE Trans. ASSP*, Vol.38(12), pp.2033-2045, 1990.

[2] V. Digalakis and H. Murveit, "GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", *Proc. ICASSP-94*, I-537, Adelaide, 1994.

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

[4] K.Shinoda, K.Iso, and T.Watanabe, "Speaker Adaptation for Demi-Syllable Based Continuous Density HMM," *Proc. of ICASSP-91*,, pp.857-860, Toronto, 1991.

[5] T. Watanabe, K. Shinoda, K. Takagi, E. Yamada, "Speech Recognition Using Tree-Structured Probability Density Function," *Proc. of ICSLP-94*, pp.223-226, Yokohama, 1994.

[6] K.Yoshida, T.Watanabe, and S.Koga, "Large Vocabulary Word Recognition Based on Demi-Syllable Hidden Markov Model Using Small Amount of Training Data," *Proc. ICASSP-89*, pp.1-4, Glasgow, 1989.

[7] C.-H.Lee, C.-H.Lin, and B.-H.Juang, "A Study on Speaker Adaptation of Continuous Density HMM parameters," *Proc. ICASSP-90*, pp.145-148, Albuquerque, 1990.