FOUR-LEVEL TIED-STRUCTURE FOR EFFICIENT REPRESENTATION OF ACOUSTIC MODELING

Satoshi TAKAHASHI *and* *Shigeki SAGAYAMA*

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

ABSTRACT

One of the problems with context-dependent HMMs is that a large number of model parameters should be estimated using a limited amount of training data. Parameters that have the same property should be tied in order to represent acoustic models efficiently. This paper proposes four-level tied-structure for phoneme models. The four levels include 1) model level, 2) state level, 3) distribution level, and 4) feature parameter level. Although some techniques have been proposed for the first three levels, feature parameter tying in the fourth level is newly proposed in this paper. We found that feature parameter tying makes it possible to represent 1,600 mean vectors of multivariate Gaussian mixture HMMs by using the combination of 16 representative mean values in each dimension. Experimental results show that feature parameter tying reduces the amount of calculation required for recognition without significant degrading performance. Furthermore, we found that feature parameter tying is also effective for model training.

1. INTRODUCTION

The HMM technique, which is a statistical method, trades model complexity off against recognition robustness. Although model complexity can be increased by using a large number of parameters, this makes it difficult to estimate the parameters accurately. If the amount of training data is not enough to estimate the parameters properly, the recognition performance degrades dramatically even if the test data is only slightly different from the training data. Contrary, if the model has an insufficient number of parameters to represent data properties, good recognition performance can not be expected.

Overviewing the past progress in acoustic model designing, the tied-structure has been one of important issues as seen in allophone tying (e.g., generalized triphone [2]), state tying (e.g., HMnet [3]), and distribution tying (e.g., tied-mixture [5] or semicontinuous HMM [6]). To reduce the total number of model parameters needed, the tied-structure is indispensable. The main problem to be pursued by the tied-structure is how to obtain a general model that represents data properties with the least complexity. There are two advantages for introducing the tied-structure. One is improved training efficiency since parameter tying reduces the total number of parameters which increases the training samples per parameter. By using the same amount of training data, the tied-structure makes it possible to obtain a model with better performance and higher robustness compared to non-tied models. Another advantage is reduced calculation amount needed for recognition through the reduction of the total number of model parameters.

This paper proposes the four-level tied-structure for the phoneme HMM. The model uses three-level tying techniques which have already been proposed, and extends the tied-structure to the feature parameter level as the fourth level. In the feature parameter tying, mean values (scalar quantity) are tied in each dimension of the mean vector of the distribution. One can say that four levels achieve tying of all constitutional elements of the HMM. No existing models realize four-level tying simultaneously.

2. FOUR-LEVEL TIED-STRUCTURE HMM

2. 1 Model level

This section describes four-level tied-structure from the top level to the bottom. The top level is tying of the allophone environments. The left and right contextual environments having the same effect on the center phoneme can share the same acoustic model. This has been realized in most existing context-dependent HMMs. We consider that the clustering of the contextual environment [1], executed in the generalized-triphone HMM [2] for example, as the method for finding the tied-structure of the contextual environment.

2. 2 State level

The second level is state tying. The states having similar feature distributions are tied across different models. State tying enables us to generate context-dependent models with a smaller number of HMM states. Basically, two strategies have been proposed to obtain the tied-state structure: the state splitting method [3] and the state merging method [4]. The state splitting method starts from a single state with a single distribution, and splits the state iteratively according to the variance of the distribution in the state. The state merging method first generates all context-dependent models that can be obtained from the training

0-7803-2431-5/95 $4.00 © 1995 IEEE

data, and similar states are merged to reduce the redundancy. Although those approaches are quite different, both methods will finally generate a similar state network that represents context-dependent models using different state paths.

2. 3 Distribution level

In the third level, similar Gaussian mixtures (having similar mean vectors and covariance matrices) are tied across different states. This basic idea is well known as the tied-mixture or the semicontinuous HMMs [5]-[7]. There are two major methods for tying the distributions. One is that after all models are trained independently, similar distributions are partially merged to make the tied-structure among all distributions. The other is that a set of distributions, which is commonly defined in all states, is prepared beforehand (like a VQ codebook), distributions related to each state are trained. In this paper, we take the former approach. All distributions in all states are clustered in order to find the tied-structure. The merged distribution of a cluster is regarded as the representative one, and the distributions in the cluster share this distribution. Distribution tying enables us to cover the feature space efficiently with a smaller number of distributions.

2. 4 Feature parameter level

Although the three tying levels described so far have already been proposed, feature parameter tying, the fourth level, is newly developed as an extension of the tying level. In feature parameter tying, the mean values are merged into some representative mean values in each dimension by using the clustering technique. The clustered mean values are tied to represent the mean vectors of the distributions.

First of all, we explain the possibility of tying mean values. For simplicity, consider the two-dimensional feature distributions whose mean vectors are μ_1 and μ_2 indicated in Figure 1. The Euclidean distance between two vectors is large since elements in dimension 1 ($\mu_{1,1}$ and $\mu_{2,1}$) are far from each other. Thus, the two vectors can not be merged in the third level. However, since elements in dimension 2 ($\mu_{1,2}$ and $\mu_{2,2}$) are close, these can be tied in the feature parameter level.

Generally speaking, context-dependent HMMs tend to contain a large number of Gaussian distributions (e.g. more than 1,000) and the same number of mean values exist in each dimension. However, when the spectral sensitivity of the feature parameter such as cepstrum is considered, less than 100 points may be enough for each dimension. Even if the mean values are merged into m points in each dimension, there still is the potential to represent m^p vectors (p is an order), and mean vectors of Gaussians will be represented by one of these vectors.

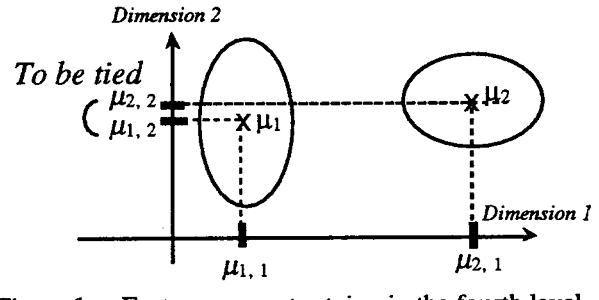

Figure 1. Feature parameter tying in the fourth level.

Next, consider the advantage from the view point of calculation cost. The log likelihood for the kth mixture component is calculated as follows when using the diagonal covariance matrix,

$$log\ p_k(\mathbf{x}_t) = -\frac{p}{2} log\ 2\pi - \frac{1}{2}\sum_{i=1}^{p} log\ \sigma_{k,i}^2 - \sum_{i=1}^{p} \frac{(x_{t,i}-\mu_{k,i})^2}{2\sigma_{k,i}^2} \quad ...... (1)$$

where $\mathbf{x}_t = (x_{t,1}, x_{t,2}, \cdots x_{t,p})^T$ (p is the order) is an input feature vector at time t, $\mu_{k,i}$ and $\sigma_{k,i}$ are the mean value and variance for the kth mixture component, respectively. When the mean values are tied across different models, the calculation for the numerator in the third term, $(x_{t,i}-\mu_{k,i})^2$, can be shared. The results are stored in a table so as to avoid recomputation in different models, thereby reducing the total calculation amount.

Furthermore, four-level tying is advantageous in model training, for example in speaker adaptation where the number of training samples available is limited. Even when a part of models are trained and their mean vectors are adapted, other mean vectors, whose elements are tied with the trained means, will be also adapted.

3. PROCEDURE FOR GENERATING FOUR-LEVEL TIED-STRUCTURE

A procedure for generating the four-level tied-structure model is described below, and the flow is also shown in Figure 2. Figure 3 schematically shows the hierarchical tying structure of the model.

[Step - 1] To construct first and second level tying simultaneously, we adopt the Successive State Splitting (SSS) method [3]. The SSS method generates a state network by iterative splitting in the contextual domain and the temporal domain to maximize the total likelihood. The network generated using this algorithm is called HMnet. Using a large amount of data from one speaker, a 600-state single Gaussian mixture HMnet was generated. The HMnet generated in the following experiments included about 1,700 triphone models.

[Step - 2] Four HMnets were cloned from the model obtained in Step-1. These were trained independently using the speech data from four different speakers. The distributions of the corresponding states in the four HMnets overlap, and a four-mixture HMnet was obtained[8]. This model was used as an initial model of the speaker-independent HMnet, and the data from 16 speakers were used to train the model.

[Step - 3] A total of 2,400 distributions in all states (4 mixtures x 600 states = 2,400) were clustered into 1,600 distributions to share the distributions. The

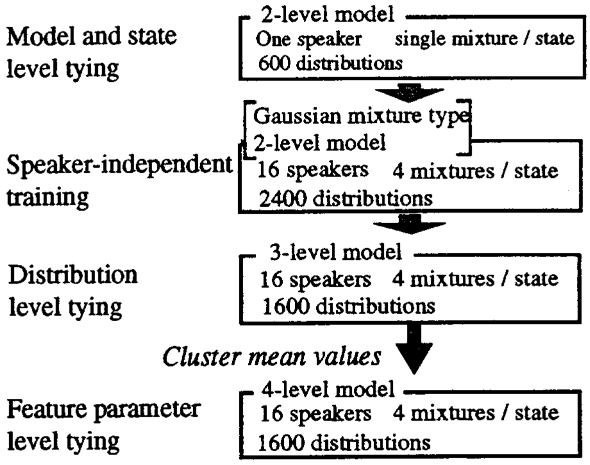
Model and state level tying
2-level model
One speaker single mixture / state
600 distributions
Gaussian mixture type
2-level model
Speaker-independent training
16 speakers 4 mixtures / state
2400 distributions
Distribution level tying
3-level model
16 speakers 4 mixtures / state
1600 distributions
Cluster mean values
Feature parameter level tying
4-level model
16 speakers 4 mixtures / state
1600 distributions

Figure 2. Generating the four-level tied-structure HMM (600 states).

Kullback-divergence was used as a metric for clustering. All distributions in cluster m were replaced by a merged distribution with mean $\mu_{m,i}$ and variance $\sigma^2_{m,i}$ calculated as follows in each dimension i,

$$\mu_{m,i} = \Big(\sum_{k \in \text{cluster } m} \mu_{k,i} \Big) / K \quad , \qquad (2)$$

$$\sigma^2_{m,i} = \Big(\sum_{k \in \text{cluster } m} \sigma^2_{k,i} + \sum_{k \in \text{cluster } m} \mu^2_{k,i} - K \cdot \mu^2_{m,i} \Big) / K \, , \qquad \ldots.. \quad (3)$$

where summations are for the distributions that belong to cluster m (K distributions in total). Each distribution tying structure remains a state with 4 Gaussian mixtures.

[Step - 4] Finally, based on the three-level model having 1,600 mean vectors, mean values were clustered into n (= 256, 64, 16, and 4) in each dimension using the scalar quantization technique, and the original mean vectors were represented by using n representative mean values. The Euclidean distance was used as a metric for clustering. Note that the covariance matrix was not changed during the experiments. The quantization method used in the third and fourth levels was the k-means clustering method.

4. EXPERIMENTS

4. 1 Baseline conditions

Models with different tying levels were compared using the 26 Japanese phoneme recognition task, a speaker adaptation task, and a word recognition task from the view point of performance, training efficiency, and the amount of calculation needed for recognition, respectively. The database used in the following experiments contained 5240-important-Japanese-word sets and 216-phoneme-balanced-word sets uttered by 20 speakers (10 males and 10 females). The even-numbered words in the 5240-word set and the 216-word set of 16 speakers (45,376 words in total) were used for training, and the odd-numbered words in the 5240-word set of the other 4 speakers were used for evaluation. The feature vector had 16 cepstrum coefficients, 16 delta coefficients, and delta power.

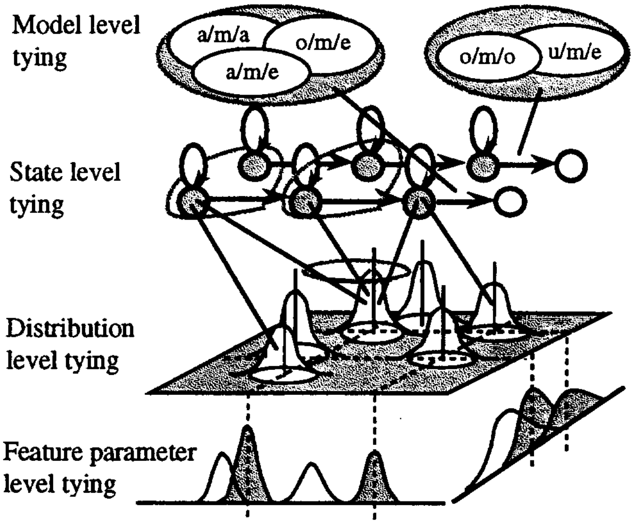

Figure 3. Hierarchical tying structure of the model.

4. 2 Results

Table 1 compares the phoneme recognition performance of models with different levels. The context-independent HMM consisting of 3 states with 16 mixtures, which performed best in different mixtures, was also tested as a reference. The performance of the two-level model was 3.2% better than that of the context-independent HMM. Although the three-level model was slightly inferior to the two-level model, the total number of parameters was decreased by 33% (from 2,400 to 1,600). For the four-level model, the number of mean clusters were varied from 256 to 4. It should be noted that the recognition performance did not degrade even with 16 representative mean values. The total number of parameters will be further decreased, if the number of clusters is determined depending on the parameter distribution of the cepstrum in each dimension.

To check the training efficiency of the four-level tied-structure model, the speaker adaptation task was performed

Table 1. 26 phoneme recognition rates.

Model	*Number of*				Average recognition rate [%]
	States	Mixtures	Distributions	Mean clusters	
Context-Independent	26x 3	16	1248	1248	84.4
2-level model	600	4	2400	2400	87.6
3-level model			1600	1600	86.8
4-level model				256	86.9
				64	86.9
				16	86.6
				4	84.0

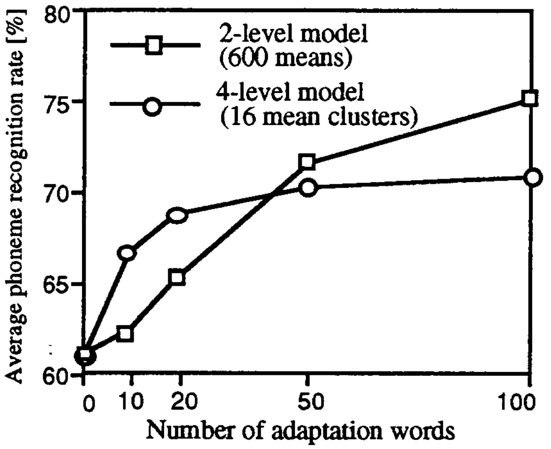

Figure 4. Average phoneme recognition rate as a function of number of adaptation words

using the two-level model with non-tied means (600 means) and the four-level model with tied means (16 means). The model of a male standard speaker was adapted for a new speaker (one-to-one speaker adaptation). A speaker-dependent HMnet (600 states, single Gaussian) was generated for the standard male speaker. The model was adapted to four target speakers (two males and two females) using maximum likelihood estimation. Figure 4 shows the average phoneme recognition performance as a function of the number of adaptation words. Because each representative mean was shared by many mean vectors (600 means / 16 clusters ≈ 38 tyings in average per representative mean), when one mean moved, many vectors were adapted simultaneously. This function worked effectively for small numbers of adaptation words. However, the performance saturates with more than 50 words due to the lack of parameter freedom. This curve will be changed when some tied means are untied to increase the parameter freedom, thereby closing to the curve for the two-level model. One effective way to use this property is that the tightly tied structure is used at the beginning of the training, and the tied means are gradually untied as the training proceeds.

The two-level model (containing 2,400 mean vectors) and the four-level model (containing 1,600 mean vectors represented by 16 means per dimension) were compared in terms of word recognition accuracy and the computation amount needed for recognition. The vocabulary size was 1,000 arbitrary selections from the odd-numbered 5240-word set, and the 100-word test sets uttered by four testing speakers were used. We counted the number of times the numerator of the third term in Eq. (1) was calculated during recognition. From the result listed in Table 2, we confirmed that feature parameter tying significantly reduces the computational requirements relating to mean values drastically without a significant loss of accuracy. This reduction is expected to be advantageous in applications on business computers, although the computation time was not reduced significantly on a workstation equipped with highly sophisticated pipeline floating arithmetics.

Table 2. Average word recognition rates and ratio for the occurrence of the calculation $(x_{t,i}-\mu_{k,i})^2$.

Model	Occurrence of calculation of $(x_{t,i}-\mu_{k,i})^2$	Word recognition rate [%]
2-level model (2400 means)	1.0	94.0
4-level model (16 mean clusters)	0.009	93.5

Note: the ratio takes account of only the calculation $(x_{t,i}-\mu_{k,i})^2$ during recognition.

5. CONCLUSION

The four-level tied-structure for phoneme HMMs was presented to represent parameters efficiently. The four levels are the model level, state level, distribution level, and feature parameter level. Feature parameter tying (tied means) is newly proposed. An exciting indication of the new technique's power is that it represented 1,600 mean vectors of Gaussians by using 16 representative mean values in each dimension, without degrading recognition performance.

A speaker-adaptation experiment confirmed that the four-level tied-structure increase training efficiency given appropriate parameter freedom. We also confirmed that feature parameter tying has the potential to reduce the computation cost of the continuous Gaussian mixture HMM although this is for just one part of the likelihood calculation.

References

[1] Sagayama S., "Phoneme Environment Clustering for Speech Recognition", Proc. ICASSP89, pp. 397-400, 1989.

[2] Lee K-F., "Context-Dependent Phonetic Hidden Markov Models for Speaker Independent Continuous Speech Recognition", IEEE Trans. ASSP, Vol 38, No 4, pp. 599-609, 1990.

[3] Takami J., Sagayama S., "A successive State Splitting Algorithm for Efficient Allophone Modeling", Proc. ICASSP92, pp. 573-576, 1992.

[4] Young S. J., Woodland P. C., "The Use of State Tying in Continuous Speech Recognition", Proc. Eurospeech93, pp. 2203-2206, 1993.

[5] Paul D. B., "The Lincoln Tied-Mixture HMM Continuous Speech Recognizer", Proc. ICASSP91, pp. 329-332.

[6] Huang X, "Phoneme Classification Using Semicontinuous Hidden Markov Models", IEEE Trans. ASSP, Vol 40, No 5, pp. 1062-1067, 1992.

[7] Hwang M-Y, Huang X, "Shared-Distribution Hidden Markov Models for Speech Recognition", IEEE Trans. ASSP, Vol 1, No 4, pp. 414-420, 1993.

[8] Kosaka T., Takami J., Sagayama S., "Rapid Speaker Adaptation Using Speaker-Mixture Allophone Models Applied to Speaker-Independent Speech Recognition", Proc. ICASSP93, pp. II-570-573.