EFFICIENT MIXTURE GAUSSIAN SYNTHESIS FOR DECISION TREE BASED STATE TYING

Tsuneo Kato, Shingo Kuroiwa, Tohru Shimizu, and Norio Higuchi

KDD R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka-shi, Saitama, 356-8502, Japan
e-mail: tkato@kddlabs.co.jp

ABSTRACT

We propose an efficient mixture Gaussian synthesis method for decision tree based state tying that produces better context-dependent models in a short period of training time. This method makes it possible to handle mixture Gaussian HMMs in decision tree based state tying algorithm, and provides higher recognition performance compared to the conventional HMM training procedure using decision tree based state tying on single Gaussian HMMs. This method also reduces the steps of HMM training procedure because the mixture incrementing process is not necessary. We applied this method to training of telephone speech triphones, and evaluated its effect on Japanese phonetically balanced sentence tasks. Our method achieved a 1 to 2 point improvement in phoneme accuracy and a 67% reduction in training time.

1. INTRODUCTION

In creating context-dependent (CD) models such as triphones, the number of possible models reaches tens of thousands and there are always some models with few or no corresponding samples in the finite training data. Parameter tying at the model or state level is essential to robustly estimate the parameters of these rarely-seen and unseen models. Decision tree based state tying[1, 2] is a top-down clustering algorithm which provides mappings to a tied state for all possible models including unseen ones. This is a widely used method and many studies on it have been reported. For example, criteria for obtaining the optimal size of state tyings were proposed in [3, 4], and approaches to generating proper binary questions for the clustering were proposed in [5, 6]. Decision tree based state tying was combined with Viterbi alignment of training data and segmental clustering to improve robustness of the models in [7].

ASR engines commonly use mixture Gaussian CD models to achieve higher performance. In the previous studies described above however, decision tree based state tying is processed on single Gaussian HMMs. As a result, the conventional method requires repetitive mixture incrementing and embedded training after the decision tree based state tying. This procedure has two disadvantages: the state tyings are produced with single Gaussian HMMs that represent acoustic characteristics of phone units very poorly, and the training procedure takes a significant amount of time due to repetitive mixture incrementing and embedded training as the number of mixtures increases. To solve these problems, we propose an effective approach to handling mixture Gaussian HMMs in decision tree based state tying. Our method produces adequate state tyings for mixture Gaussian CD models because clustering is processed on the same number of mixture Gaussians as the target models used for speech recognition. Furthermore, this method greatly shortens the training time by skipping a large part of the mixture incrementing and embedded training steps. The proposed method synthesizes mixture Gaussian distributions in the clustering process and tied-state CD models of mixture Gaussian HMMs are output as the results.

In Section 2, we review conventional decision tree based state tying. Section 3 points out the insufficiencies due to the limits of single Gaussian in the conventional method and then describes our approach to handling mixture Gaussian HMMs. Experimental results on Japanese phonetically balanced sentence tasks are presented in Section 4. Finally, Section 5 presents our conclusions.

2. DECISION TREE BASED STATE TYING

Decision tree based state tying is a top-down clustering process. Assuming that triphones are used as CD models, all HMM states for corresponding positions of triphones derived from a monophone are collected as a root node at the beginning . Starting from the root node, nodes are successively split into two successor nodes and a tree like that shown in Fig. 1 is produced by clustering. The leaf nodes are treated as equivalent classes where all states are tied to one state. When a node is split into two successor nodes, one of predefined binary questions related to phonetic contexts is chosen according to a criterion of goodness-of-split maximization. A change in log likelihood for outputting corresponding frames in training data is often used as the goodness-of-split. The clustering proceeds until stopping conditions are fulfilled. The stopping conditions are as follows: the goodness-of-best-split falls below a threshold or the number of frames that states in the node occupies in the training data falls below a threshold.

Let node S_m be split by question q into two successor nodes, $S_{m,y}(q)$ and $S_{m,n}(q)$. The change in log likelihood is given by

$$\Delta L_q = L(S_{m,y}(q)) + L(S_{m,n}(q)) - L(S_m) \qquad (1)$$

The log likelihood $L(S_m)$ is calculated by an approximate

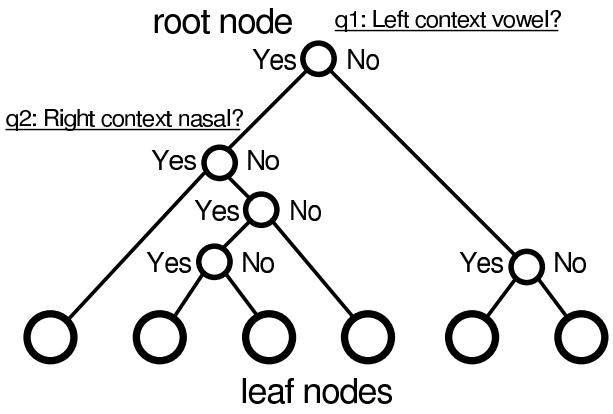


Fig. 1. Phonetic decision tree

function of the variance of feature vectors and the expected number of occupied frames in the training data, since the direct calculation of log likelihood for all these frames would otherwise take an enormous amount of time. For this approximation, the mean vector and diagonal covariance matrix of node S_m are given by the following equations.

$$\hat{\mu}_m^{(k)} = \sum_i \Gamma_{m,i}\mu_{m,i}^{(k)} \Big/ \sum_i \Gamma_{m,i} \tag{2}$$

$$\begin{aligned}\hat{\sigma}_m^{(k)} &= \left[\sum_i \Gamma_{m,i}(\mu_{m,i}^{(k)} - \hat{\mu}_m^{(k)})^2 \right. \\ &\quad \left. + \sum_i \Gamma_{m,i}\sigma_{m,i}^{(k)}\right] \Big/ \sum_i \Gamma_{m,i} \end{aligned} \tag{3}$$

where $\mu_{m,i}$, $\sigma_{m,i}$ and $\Gamma_{m,i}$ denote the mean vector, diagonal covariance matrix and expected number of occupied frames by i'th state in node S_m. The number k denotes the k'th component of feature vector. Given an observation sequence in the training data $O_t(t = 1, \ldots, T)$, the log likelihood for node S_m is given by

$$\begin{aligned} L(S_m) &\approx \sum_{t=1}^{T} \log[N(O_t, \hat{\mu}_m, \hat{\sigma}_m)] \cdot \gamma_t(m) \\ &= -\frac{1}{2}(K\log 2\pi + \log|\hat{\sigma}_m| + K)\Gamma_m \end{aligned} \tag{4}$$

where $\gamma_t(m)$ and Γ_m denote the probability that states of node S_m are occupied at time t, and the expected number of occupied frames in the sequence. Here, $N(O_t, \hat{\mu}_m, \hat{\sigma}_m)$ indicates the probability that a Gaussian distribution of the mean vector $\hat{\mu}_m$ and covariance matrix $\hat{\sigma}_m$ will output an observation O_t.

The goodness-of-split is thus found by the approximation for producing a decision tree. After the clustering process, tied-state triphones of single Gaussians with the largest variance in leaf nodes, or with the mean vector $\hat{\mu}_m$ and covariance matrixs $\hat{\sigma}_m$.

Since the decision tree based state tying outputs single Gaussian HMMs, mixture Gaussian HMMs actually used

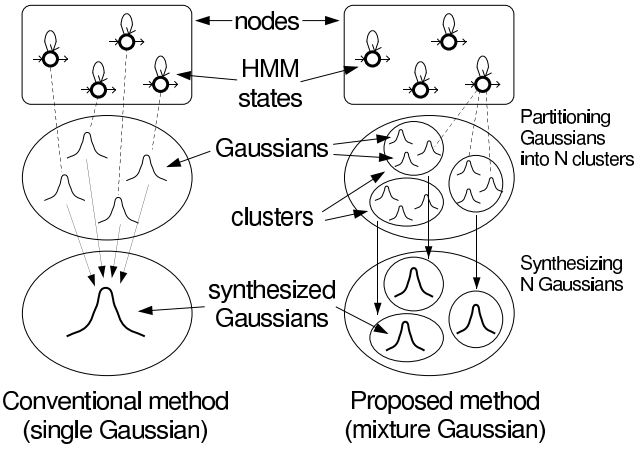


Fig. 2. Synthesis of mixture Gaussian distributions for a node in the proposed method in comparison with the conventional method

in speech recognition are obtained through repetitive mixture incrementing and embedded training after the decision tree based state tying. Mixture incrementing doubles the number of Gaussians, by copying a distribution with a perturbed mean vector and providing half of the original mixture weight to both distributions. These steps are repeated until HMMs of the desired number of mixtures or the required performance are obtained.

3. SYNTHESIS OF MIXTURE GAUSSIANS FOR DECISION TREE BASED STATE TYING

In equation (4), the log likelihood $L(S_m)$ which remotely determines the state tyings is calculated on a single Gaussian distribution. However as their superior performance in speech recognition shows, mixture Gaussian distributions represent acoustic characteristics much more precisely than single Gaussians. We therefore conclude that the decision tree based state tying should be processed on mixture Gaussians. We also conclude that the decision tree based state tying on distributions with the same resolution as the target mixture Gaussian models produces better state tyings.

Our proposed method handles mixture Gaussian HMMs for both input untied triphones and output tied-state triphones, and also assumes mixture Gaussian expressions during the clustering process. At every node of the decision tree, all Gaussian components consisting of states in the node are partitioned into N classes, where N indicates the maximal number of mixtures among the states in the node. N mixture Gaussian distributions are then newly synthesized for the node. Fig. 2 shows differences between the conventional and proposed methods.

Partitioning of Gaussian components is carried out by K-means clustering algorithm. The mean vectors of input models are the elements and variance-scaled Euclidean distance is the distance metric of the clustering. Following the partitioning of Gaussian components, a new Gaussian distribution is synthesized for each class of Gaussian components. The mean vector and diagonal covariance of n'th

newly synthesized Gaussian distribution and its mixture weight are given by the following equations.

$$\hat{\mu}_{m,n}^{(k)} = \sum_i \Gamma_{m,n,i} \mu_{m,n,i}^{(k)} \Big/ \sum_i \Gamma_{m,n,i} \tag{5}$$

$$\begin{aligned} \hat{\sigma}_{m,n}^{(k)} &= \left[\sum_i \Gamma_{m,n,i} (\mu_{m,n,i}^{(k)} - \hat{\mu}_{m,n}^{(k)})^2 \right. \\ &\quad \left. + \sum_i \Gamma_{m,n,i} \sigma_{m,n,i}^{(k)} \right] \Big/ \sum_i \Gamma_{m,n,i} \end{aligned} \tag{6}$$

$$w_{m,n} = \sum_i \Gamma_{m,n,i} \Big/ \sum_n \sum_i \Gamma_{m,n,i} \tag{7}$$

The expected number of occupied frames by i'th component in n'th classes $Gamma_{m,n,i}$ is approximated by the product of expected number of occupied frames by the original state and the component's mixture weight.

After synthesizing the N mixture Gaussian distributions, They are used for calculation of log likelihood for the training data. If we took overlaps between Gaussian distributions into account, the integral for the overlaps would require an immense amount of computations. Hence in our proposed method, we neglect the overlaps. We also approximate the log likelihood derived from the mixture Gaussian distributions by the maximum value among them. Instead of equation (4), the log likelihood for the node S_m is approximated by

$$\begin{aligned} & L(S_m) \\ &\approx \sum_{t=1}^{T} \log \left[\sum_{n=1}^{N} w_{m,n} N(O_t, \hat{\mu}_{m,n}, \hat{\sigma}_{m,n}) \right] \cdot \gamma_t(m) \\ &\approx \sum_{t=1}^{T} \log \left[\max \{ w_{m,n} N(O_t, \hat{\mu}_{m,n}, \hat{\sigma}_{m,n}) \} \right] \cdot \gamma_t(m) \\ &\approx \sum_{n=1}^{N} [\Gamma_{m,n} \log w_{m,n} - \frac{\Gamma_{m,n}}{2} (K \log 2\pi + \log |\hat{\sigma}_{m,n}| + K)] \\ &= \sum_{n=1}^{N} \left[\Gamma_{m,n} \log \Gamma_{m,n} - \frac{\Gamma_{m,n}}{2} (K \log 2\pi + K \right. \\ &\quad \left. + \log |\hat{\sigma}_{m,n}|) \right] - \sum_{n=1}^{N} \Gamma_{m,n} \cdot \log \sum_{n=1}^{N} \Gamma_{m,n} \end{aligned} \tag{8}$$

where, $\hat{\mu}_{m,n}$, $\hat{\sigma}_{m,n}$, $w_{m,n}$, $\Gamma_{m,n}$ denote the mean vector, covariance matrix, mixture weight, and expected number of occupied frames of the newly synthesized n'th Gaussian distribution for the node S_m. The flow chart of our proposed method is presented in Fig.3.

4. EXPERIMENTAL RESULTS

The performance of telephone speech triphones trained by the proposed and conventional methods were compared in the recognition of phonetically balanced sentences.

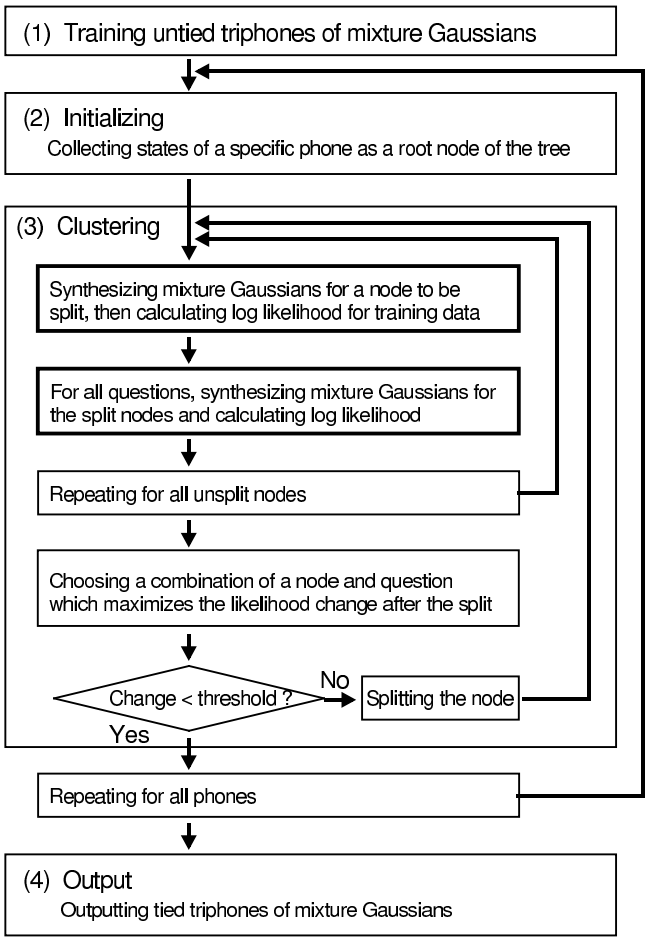


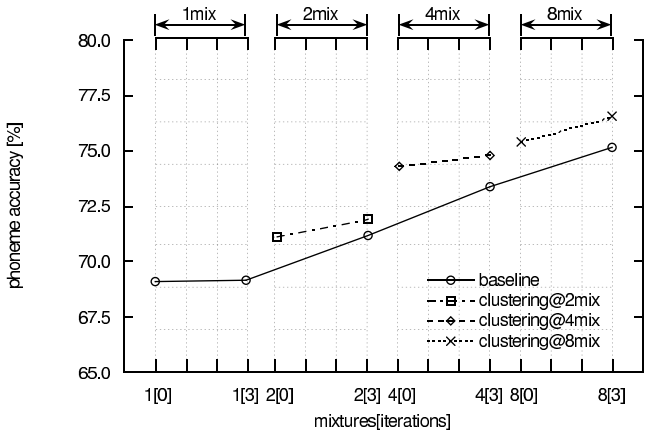
Fig. 3. Flow chart of state tying.

The conventional HMM training procedure was comprised of the following steps.

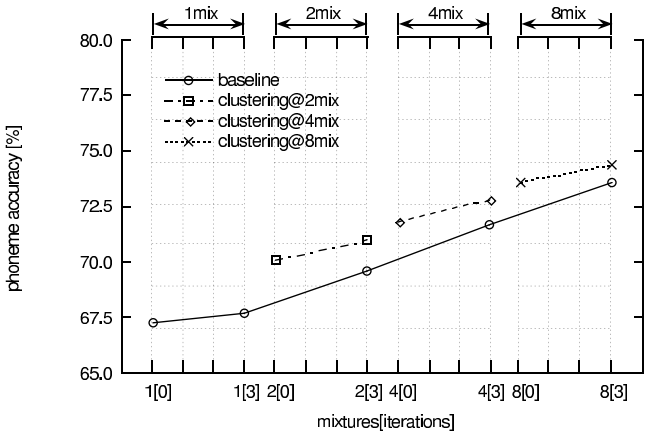
1. Training untied triphones of single Gaussians
2. Processing conventional decision tree based state tying on the triphones
3. Repeating the embedded training three times
4. Mixture incrementing to obtain 2 mixture triphones
5. Repeating the embedded training three times
6. Mixture incrementing to obtain 4 mixture triphones
7. Repeating the embedded training three times
8. Mixture incrementing to obtain 8 mixture triphones
9. Repeating the embedded training three times

Triphones of two, four, and eight mixtures were obtained in the course of these steps. Our proposed method onn the other hand was comprised of the following steps.

1. Training untied triphones of mixture Gaussians
2. Processing proposed decision tree based state tying on the triphones of mixture Gaussians
3. Repeating the embedded training three times



(a) Male models



(b) Female models

Fig. 4. Phoneme accuracy as a function of the number of mixtures and iterations of embedded training.

Triphones of two, four, and eight mixtures were obtained in separate procedures. Evaluation was made on triphones just after the decision tree based state tying and on triphones after three iterations of embedded training.

The two methods were applied for training male and female gender-dependent models. Training data were 9,000 phonetically balanced sentences uttered by 1,057 male speakers and 10,700 sentences uttered by 506 female speakers. Feature parameters were 12 MFCC coefficients, their first and second order derivatives, and the first and second order logarithmic power derivatives. The triphones after decision tree based state tying had nearly the same number of HMM states (1200). Male models and female models were evaluated separately. Test sets were 990 phonetically balanced sentences uttered by 30 male and 30 female speakers.

Accuracy of both male and female models for phoneme recognition with Japanese syllabic constraints are shown on Fig. 4. The horizontal axis indicates the number of mixtures of a state and the figures in the brackets are the number of embedded training iterations. The dashed lines represent the scores of our proposed method for two, four, and eight mixtures, whereas the solid line represents scores of the conventional method. First of all, it is clear that our proposed method reduced the number of HMM training steps, for example from 12 iterations of embedded training to 3 iterations in the 8 mixture case. As a result, our proposed method achieved a 67% reduction in the training time despite increased computations in the decision tree based state tying. In terms of phoneme accuracy, the triphones just after our proposed decision tree based state tying exceeded those obtained by mixture incrementing. Even after three iterations of embedded training, the triphones of our proposed method exceeded those of the conventional method, being 1 to 2 points higher in all the cases for both male and female models. (Scores after three iterations of embedded training were found to be nearly saturated to the maximum score.) Better initial models for embedded training maintains their superior scores after several iterations of embedded training.

5. CONCLUSIONS

This paper has presented an effective approach to synthesizing mixture Gaussian distributions in decision tree based state tying. This method makes it possible to handle mixture Gaussian HMMs with decision tree based state tying algorithm and provides better state tyings for target mixture Gaussian HMMs used in speech recognition. We applied this method to training of telephone speech triphones. Experimental results on phonetically balanced sentence tasks showed a 1 to 2 point improvement in phoneme accuracy. Our method also greatly reduced the steps of HMM training procedure and achieved a 67% reduction in training time.

6. REFERENCES

[1] L. R. Bahl et al. "Decision trees for phonological rules in continuous speech". In Proc. ICASSP 91, pages 185–188, 1991.

[2] S. J. Young, J. J. Odell, and P. C. Woodland. "Tree based state tying for high accuracy acoustic modeling". In ARPA Workshop on Human Language Technology, pages 286–291, 1994.

[3] K. Shinoda and T. Watanabe. "Acoustic modeling based on the MDL principle for speech recognition". In EuroSpeech 97, pages 99–102, 1997.

[4] W. Chou and W. Reichl. "Decision tree state tying based on penalized bayesian information criterion". In Proc. ICASSP 99, pages 2481–2484, 1999.

[5] R. Singh, B. Raj, and R. M. Stern. "Automatic clustering and generation of contextual questions for tied states in hidden markov models". In Proc. ICASSP 99, pages 117–120, 1999.

[6] D. Willett et al. "Refining tree-based state clustering by means of formal concept analysis, balanced decision trees and automatically generated model-sets". In Proc. ICASSP 99, pages 565–568, 1999.

[7] W. Reichl and W. Chou. "Decision tree state tying based on segmental clustering for acoustic modeling". In Proc. ICASSP 98, pages 801–804, 1998.