A New State-dependent Phonetic Tied-Mixture Model with Head-Body-Tail Structured HMM for Real-time Continuous Phoneme Recognition System

Junho Park *Hanseok Ko*

ISPL, Department of Electronics Engineering, Korea University, Seoul, Korea
jhpark@ispl.korea.ac.kr hsko@korea.ac.kr

Abstract

An acoustic model for a real-time continuous phoneme recognition system must exhibit the following desirable feature: an ability to minimize the recognition performance degradation while solving the model complexity problem to confine the delay to a minimum in recognition process. To cope with the challenges, we introduce the state-dependent Phonetic Tied-Mixture (PTM) model with Head-Body-Tail (HBT) structured HMM as an acoustic model optimization. The proposed acoustic modeling method shows a significant improvement in recognition performance and becomes a solution to the sparse training data problem and the model complexity problem. Moreover, defining the exceptional Gaussian mixtures in tying process achieves a drastic reduction in phoneme error rate compared to traditional state-dependent PTM method. In this paper, we describe the new acoustic model optimization procedure and show the outstanding performance evaluation results for real-time continuous phoneme recognition system.

Index Terms: acoustic modeling, state-dependent PTM

1. Introduction

Speech recognition technology can be employed in many useful applications as the principal human-machine interface. Those applications may need just a simple recognizer for some interested vocabularies rather than a complex one, which contain infrequently used large vocabularies. The acoustic model for the simple recognizer has relatively small complexity and needs a few storage memories in the application system. For a large vocabulary tasks, however, the speech recognition system should guarantee a large sized and highly performing acoustic model for reliable recognition results. In the case of real-time phoneme recognition, it is desirable to realize a system with just a simple acoustic model though its covered vocabulary range is large. To achieve this objective, many HMM-based or SVM-based acoustic modeling methods for phoneme recognition system have been investigated. There has been a number of research efforts conducted in the SVM-based methods for speech recognitions. Simple vowel classifications using SVM by Clarkson [1], dynamic time alignment kernel method by Shimodaira [2], and SVM/HMM hybrid method by Golowich [3] are some of the work done so far for utilizing the SVM-based approach for speech recognition. HMM algorithm has often been used for such an approach. Using an algorithm for phoneme recognition based on HMM, Tamura [4] tried to find an optimum phoneme sequence for visual speech parameter from sentences in ML sense. Yamamoto [5] also proposed a novel lip movement synthesis method of mapping input speech based on HMM based phoneme recognition. In this paper, we design the acoustic model for phoneme recognition targeting a real time lip-synch system that activates 2-D avatar's lip motion in synch with incoming speech utterance. To achieve the real-time continuous phoneme recognition (or classification), we employ the alternative state-dependent PTM (Phonetic Tied-Mixture) model with HBT (Head-Body-Tail) structured HMM. Context independent (CI) model alone does not well represent continuously uttered phoneme sequences in all contexts though its size is relatively small. As a result, CI models do not achieve high recognition performance for continuous phoneme recognition tasks. On the other hand, context dependent (CD) models provide a more detailed description of the acoustic units undergoing analysis. However, the gain in the detailed acoustic models requires an adequate amount of model complexity and memory size. This problem can be, in part, resolved by employing either tied-mixture (TM) continuous parameter, phonetic tied-mixture (PTM), or subspace distribution clustering (SDC) modeling [6][7][8]. These methods, however, could not effectively represent several varying articulations of continuously uttered phoneme sequences. The proposed alternative state-dependent PTM model with Head-Body-Tail structured HMM provides itself as a solution to finding reliable model parameter values with a very simple mixture clustering algorithm.

In the next section we first describe the proposed state-dependent PTM with Head-Body-Tail structured HMM. We then prescribe in Section 3, the optimizing method of state-dependent PTM with HBT structured HMM for continuous phoneme recognition system. In Section 4, we present the representative experiments that show how

effective the proposed state-dependent PTM with HBT structured HMM is, for optimized embedded speech recognition system. Finally, concluding remarks are presented in Section 5.

2. State-dependent PTM with HBT structured HMM

The semi-continuous or tied-mixture HMM, in general, has state output probabilities compromised between discrete and continuous distribution HMM's. That is, the global codebook of the TM HMM, with Gaussian code-words, is constructed by a mixture tying process, whose state output probability is the weighted sum of these code-words. Eq. (1) shows the output probability of state *S* for input vector $\mathbf{x}_t$.

$$b_S(\mathbf{x}_t) = \sum_{l=1}^{L} b_S(l) G(\mathbf{x}_t; \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l) \tag{1}$$

where *L* is the size of the codebook, $b_S(l)$ is the weight for each codebook index *l*, and $G(\bullet)$ denotes Gaussian distribution.

The continuous distribution HMM (CDHMM) has a Gaussian mixture and corresponding weights for each state. The tied-mixture modeling method is used to make one or more codebooks for tying Gaussians from the continuous density Gaussian mixture. Fig. 1 shows the concept of tied-mixture (TM) modeling.

2.1 State-dependent PTM modeling method

The context dependent HMM must share their parameters such as the state distributions for reliable parameter estimation. The generation of the state-dependent PTM model is based on both the state tying and mixture tying for an efficient complexity reduction of triphone models [7]. Compared to the pure TM or PTM models, the state-dependent PTM model represent the state identity using small-sized several state-dependent Gaussian codebooks. The state-dependent PTM model uses a small set of parameters to discard the overlapping mixture distributions for robust model estimation.

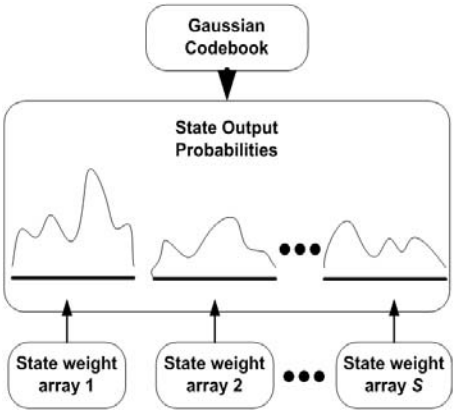

Figure 1: Concept of tied-mixture modeling method

In order to deal with the overlapping Gaussian components among the tied states, we use the Gaussian clustering based on the Bhattachyaryya distance measure. Given two Gaussian components, $G_1(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ and $G_2(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)$, the distance measure is represented as

$$D(G_1, G_2) = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \left(\frac{\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2}{2}\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \ln \frac{\left|(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2)/2\right|}{\left|\boldsymbol{\sigma}_1\right|^{1/2} \cdot \left|\boldsymbol{\sigma}_2\right|^{1/2}} \tag{2}$$

As a result, the state-dependent PTM model shares some state output distributions while its state shares common Gaussians in the state-dependent codebook. This state-dependent PTM model has properties of both the state-clustered and the tied-mixture models. Additionally, in the state clustering process, the state-dependent PTM modeling method could generate decision-trees for unseen contexts. Fig. 2 shows the structure of state-dependent PTM modeling method.

2.2 HBT structured HMM for phoneme recognition

Co-articulation effects for several contexts can be represented by context dependent HMM. One such model set, referred to as HBT structured HMM [9], has been used effectively in connected digit recognition. HBT models are a special case of subword modeling that represents the beginning, middle, and end of a word. The center of each phoneme, modeled by the body model, is a context independent unit. Context dependency information is incorporated in the head and tail models. Fig. 3 shows an example of HBT structured HMM for 'E', 'A', and 'Oh' phoneme sequence..

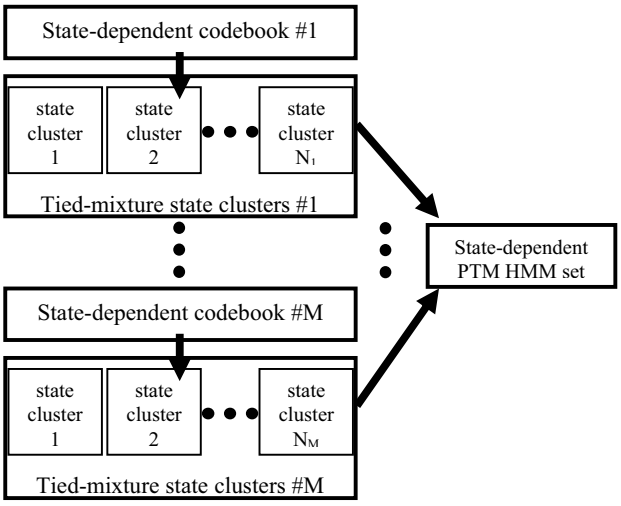

Figure 2: Structure of state-dependent PTM model

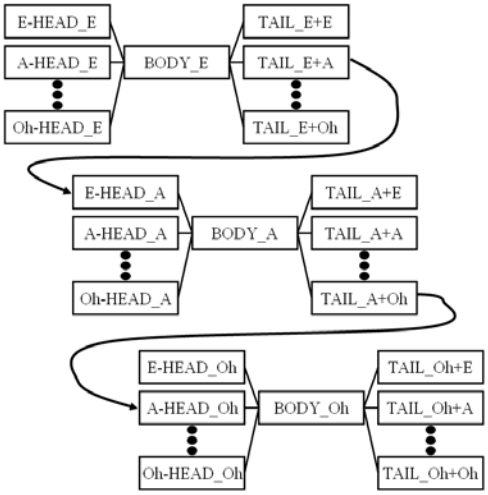

Figure 3: Example of HBT model structured HMM for 'E', 'A', and 'Oh' phoneme sequence

3. Optimizing a new state-dependent PTM with HBT structured HMM

A new state-dependent PTM with HBT structured HMM has various optimizing targets. The objectives are: to enhance the representation ability of context dependency information and to reduce the redundant Gaussians in the model. In this section, we introduce our methods for optimizing the new model.

3.1 Exceptional Gaussians in the mixture tying process

In the HBT structured HMM concept, the context dependency information, as mentioned in Section 2, is incorporated in the head and tail model. Moreover, the first state of head model and the last state of tail model has the very important transition information of co-articulation effects due to prior and posterior phonemes, respectively. In this point of view, we set the Gaussian set of the first states of head models and the last states of tail models as the exceptional Gaussians in the mixture tying process.

3.2 Variable size of state-dependent codebook

In the state-dependent PTM model, unique size of state-dependent codebook leads to the error between the likelihood of original and tied Gaussian mixtures for some models needing highly varying phonemes. On the other hand, too many small weighted mixtures in the codebook increase the model complexity. To cope with that, we adjust the codebook size of each state to minimize the likelihood lost in the Gaussian merging process. Given two Gaussian components, $G_1(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ and $G_2(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)$ with their relevant occurrence counts in the training set c_1 and c_1, the merging process based on the likelihood loss computation [10] is represented as follows.

$$c = c_1 + c_2 \tag{3}$$

$$\boldsymbol{\mu} = \frac{c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2}{c} \tag{4}$$

$$\boldsymbol{\sigma} = \frac{c_1}{c}[\boldsymbol{\sigma}_1 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^{\mathbf{T}}] + \frac{c_2}{c}[\boldsymbol{\sigma}_2 + (\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^{\mathbf{T}}] \tag{5}$$

The likelihood loss due to the merge of $G_1(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ and $G_2(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)$ is;

$$\Delta(G_1 + G_2) = \frac{c\log|\boldsymbol{\sigma}| - c_1\log|\boldsymbol{\sigma}_1| - c_2\log|\boldsymbol{\sigma}_2|}{2}. \tag{6}$$

The codebook size to minimize the likelihood loss can be determined by checking the accumulated value of Eq. (6).

3.3 Common methods for optimizing TM models [11]

In the model training procedure, we define the state weight lower boundary with $2\text{x}10^{-5}$. The lower boundary prevents the state weights from falling into zero. In our model, the largest part of a model size is the set of state weight arrays. The data type of state weight arrays is floating point. Consequently, a 4-byte memory is needed per weight value. However, if we quantize the weights into 256 levels, each weight requires only 1-byte memory. Moreover, the state level Gaussian selection algorithm uses a sub-set of pre-computed Gaussians for calculating each output probability. The Gaussian selection criterion we use is the likelihood difference of each Gaussian to the maximum likelihood valued Gaussian.

4. Experiments

In this section, we evaluate the effectiveness of the proposed acoustic model implementation for continuous phoneme recognition system using the state-dependent PTM with HBT structured HMM with optimization.

4.1 Comparison of the simple CI and HBT models

In this experiment, we compare the recognition performance between CI and HBT models that are continuous distributed HMM. The simple CI model has 9 states and the HBT model has totally 9 states (3 states per Head, Body, Tail). Both models for comparison have 8-mixture per state. The feature we used is 39th order MFCC's constructed with 12th order MFCC's, log energy, their delta, and delta-delta values.

In this test, we used a set of single vowel classes embedded in 452 phonetically balanced Korean words set both as the training model and test DB in order to simplifying the process. We merged Diphthongs into corresponding single vowels to apply our model to real-time lip-synch system. For speaker independent tasks, we precluded the test speaker's utterances from those in

training. The total number of Gaussians for CI and HBT model is 552 and 8,616, respectively. Table 1 shows both recognition performances of each model. In this experiment, both phone error rate (PER) and sentence error rate (SER) of CI model are about 2 times larger than those of HBT model. This experimental result shows that the context dependency information in head and tail model could lead drastic improvement in terms of phoneme recognition performance.

4.2 Performance evaluation of the state-dependent PTM with HBT structured HMM

In this test, we evaluated the performance of the state-dependent PTM with HBT structured HMM. Our set up of this test is same with above experiment. We, firstly, constructed continuous distributed HBT model as above test. Secondly, we developed the traditional state-dependent PTM model in HBT structured HMM. The total number of Gaussians in the model is 2,064. And finally we constructed state-dependent PTM model having the exceptional Gaussians in the mixture tying process. The total number of Gaussians in this case is 2,008. In this experiment, we set the number of Gaussians to 32 per state-dependent codebook, uniquely. As shown in Table 2, the traditional state-dependent PTM method could not reflect context dependency information of head and tail models, effectively. In other words, the mixture tying process of head and tail model could degrade the advantage of the HBT model in the aspect of context dependent modeling for co-articulation effect.

In the performance evaluation of variable state-dependent PTM codebook size, the average codebook size of our model is 32. In this experiment, as shown Table 3, the phoneme recognition performance is slightly increased though the total number of Gaussians is same with the fixed codebook sized model.

Table 1: Performance evaluations of CI and HBT model

	CI model	HBT model
PER	22.54%	9.58%
SER	77.43%	40.09%

Table 2: Performance evaluations of state-dependent PTM with HBT structured HMM

	Conventional state-dependent PTM with HBT HMM	New state-dependent PTM with HBT HMM (exceptional Gaussian)
PER	19.74%	9.21%
SER	53.63%	39.73%

Table 3: Performance evaluations of variable size of state-dependent codebook in HBT structured HMM

	Fixed codebook size	Variable codebook size
PER	9.21%	9.11%
SER	39.73%	39.61%

5. Conclusions

In this paper, we implemented an acoustic model based on a state-dependent PTM with HBT structured HMM, targeting for real-time continuous phoneme recognition applicable to real-time lip-synch system. Moreover, we investigated candidate optimization techniques, such as mixture tying process with exceptional Gaussians in head and tail model, variable state-dependent PTM codebook sized method, and others suitable for tied-mixture model. These optimization schemes made the proposed model more effective and reliable in the continuous phoneme sequence recognition. Experimental results showed that the new state-dependent PTM with HBT structured HMM can be reliably applied to real-time lip-synch system or simple vowel segmentation system.

6. Acknowledgements

This work was supported by grant NO. R01-2006-000-11162-0(2006) from the Basic Research Program Korea Science and Engineering Foundation of Ministry of Science & Technology

7. References

[1] P. Clarkson and P. Moreno, "*On the use of support vector machines for phonetic classification,*" ICASSP*1999*, pp585-588.

[2] H. Shimodaira, et al, "*Support Vector Machine with Dynamic Time-Alignment Kernel for Speech Recognition*," Eurospeech*2001*.

[3] S. Golowich and D. Sun, "*A Support Vector/Hiiden Markov Model Approach to Phoneme Recognition*," ASA Proceedings of the Statistical Computing Section, 1998, pp.125-130.

[4] M. Tamura, et al, "*Visual speech synthesis based in parameter generation from HMM": Speech driven and text-and-speech driven approaches,*" Proc. AVSP 98, International Conference on Auditory-Visual Speech Processing.

[5] E. Yamamoto, et al, "*Lip movement synthesis from speech based on Hidden Markov Models*," Speech Communication, Vol. 26, 1998, pp.105-115.

[6] S. Young, "*The general use of tying in phoneme-based HMM speech recognizers,*" ICASSP*1992, pp. 569- 572.*

[7] A. Lee, et al, "*A new phonetic tied-mixture model for efficient decoding,*" ICASSP*2000, pp. 1269-1272.*

[8] E. Bocchieri and B. Mak, "*Subspace distribution clustering hidden Markov model,"* IEEE Trans on speech and Audio Processing*, vol. 9, March 2001, pp. 264-276.*

[9] W. Chou, B. Juang, C. Lee, "*Minimum error rate training based on N-best string models*," ICASSP*1993*, pp. 652-655.

[10] L. Deng, et al, "*High-performance robust speech recognition using stereo training data*," ICASSP*2001*, pp.301-304.

[11] J. Park and H. Ko, "*Compact Acoustic Model for Embedded Implementation*," ICSLP*2004*, pp693-696, 2004.