A NEW PHONETIC TIED-MIXTURE MODEL FOR EFFICIENT DECODING

Akinobu Lee [*] *Tatsuya Kawahara* [*] *Kazuya Takeda* [†] *Kiyohiro Shikano* [‡]

[*] Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[†] Nagoya University, Nagoya 464-8603, Japan
[‡] Nara Institute of Science and Technology, Ikoma 630-0101, Japan

ABSTRACT

A phonetic tied-mixture (PTM) model for efficient large vocabulary continuous speech recognition is presented. It is synthesized from context-independent phone models with 64 mixture components per state by assigning different mixture weights according to the shared states of triphones. Mixtures are then re-estimated for optimization. The model achieves a word error rate of 7.0% at 20k-word dictation of newspaper corpus, which is comparable to the best figure by the triphone of much higher resolutions. Compared with conventional PTMs that share Gaussians by all states, the proposed model is easily trained and reliably estimated. Furthermore, the model enables the decoder to perform efficient Gaussian pruning. It is found out that computing only two out of 64 components does not cause any loss of accuracy. Several methods for the pruning are proposed and compared, and the best one reduced the computation to about 20%.

1. INTRODUCTION

Sharing and tying of HMM parameters for accurate and efficient acoustic modeling has been a major concern in large vocabulary continuous speech recognition systems. A typical triphone model has thousands of states and hundreds of thousands of Gaussian distributions. Estimation of such a large number of parameters requires huge amount of training data to obtain the desired accuracy. Thus sharing and tying of the models in various levels are widely adopted to reduce the number of total parameters.

The current dominant approach of parameter tying is the state sharing and clustering of triphone model according to acoustic similarity. Another approach is a tied-mixture (TM) system where a single set of Gaussian distributions is shared by all HMMs while each state has different mixture weights. And also there is phonetic tied mixture (PTM) HMMs[1][2], where a set of Gaussian components is defined independently for each phone and the triphone variants of the same base phone share the Gaussians set.

The TM and PTM HMMs are advantageous to the state-clustered triphone in that overlapping mixture distributions on different states are properly modeled with less Gaussians, so the training can be more reliable. But in conventional TM and PTM HMMs, all Gaussian distributions in different states (within a phone in PTM) should be covered by one codebook, and the size often gets very large to the extent of hundreds or thousands. It is not easy to train such a large mixture to an optimal point. In this paper we propose a PTM model of another parameter sharing scheme, a phonetic *state-based* tied mixture model that realizes both easy training and reliable estimation.

Another merit of TM over the triphone model is that since it has a larger codebook with less redundant distributions, it is easier to introduce pruning mechanisms in Gaussian mixture computation. Therefore, several pruning methods are also proposed and compared.

2. PHONETIC TIED-MIXTURE MODEL

A triphone model defines separate states for each context-dependent phone variants in order to represent the cross-phone articulation effects precisely. But as many context-dependent states have their own mixtures, there can be many overlapping mixtures among them. These cause the number of parameters to grow improperly and make parameter estimation unreliable.

In TM HMMs, on the other hand, a large number of mixture components are shared within all states and the states have different mixture weights for each. As the whole acoustic space is modeled with larger mixture units, the overlapping mixture components are well represented with less parameters. However, TM models have not demonstrated as good performance as the triphone model. The total number of mixture components of TMs is usually smaller by a magnitude, thus it can not have enough discriminative ability. The root cause is that it is not easy to train or estimate a large scale codebook of distributions as a whole.

Based on these viewpoints, we propose another type of PTM. By sharing a set of mixture components among states of the same topological location, redundant components in all triphone are merged. Compared with conventional PTMs, the Gaussian distributions is modeled more accurately by having independent mixture components on different topological location.

The proposed PTM HMM is synthesized from state-clustered triphones and a monophone model as illustrated in Figure 1. The mixtures of monophone HMMs are assigned to the corresponding states of the tied-state triphone HMMs, where each non-shared state in triphones shares only the mixture components and have different mixture weights. After the assignment, the overall mixtures and weights are re-estimated for optimization.

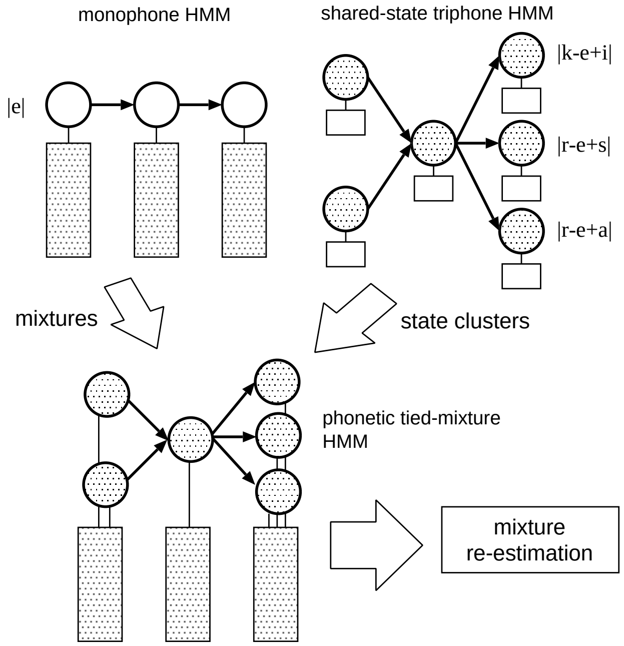


Figure 1: Training of proposed PTM HMMs

The construction process is straight-forward and easy. Mixtures are estimated by monophone models by gradually increasing their numbers of components. This makes it easy to reliably estimate large number of mixtures. Context-dependent modeling is realized by the state-clusters of the conventional triphones. Here, corresponding mixture components are substantially assigned by weighting among a large codebook that efficiently covers the whole acoustic space. The actual steps for building this model is as follows:

1. Train monophone HMMs that have a large Gaussian mixture for each state. No parameter tying is done here.

2. Train shared-state triphone HMMs using the same phone set. The mixture components on this model is used only for determining state-sharing, so a single Gaussian is enough. Any method of can be used to determine the state sharing, but the number of states in a phone must be the same as the monophone on step 1, and only states of the same topological location within a phone are allowed to share.

3. Assign the mixtures of monophone HMMs to corresponding states of the triphone HMMs. Tied triphone states also share both the same mixture components and weights, and non-tied states not shared on step 2 share only the mixture components and have different mixture weights.

4. Re-estimate the mixture components and assigned weights to discriminate triphones. The mixture itself is also re-trained.

3. GAUSSIAN PRUNING METHODS

Acoustic matching often occupies the largest part of processing time in current recognition systems, because a large amount of Gaussian distributions must be computed. So reducing the cost is significant for fast decoding.

Gaussian pruning is one approach to reduce the amount of computed Gaussians. As the log likelihood of a Gaussian density keeps descending while accumulating distances for each vector element, we can determine whether the value of the Gaussian will be below a certain threshold before computing all the distance components. Especially, since PTM HMMs have a larger mixture codebook, such a pruning mechanism works more effectively than tied-state triphone HMMs.

We propose several Gaussian pruning methods for reduction of computational cost. The purpose is to get k-best Gaussians out of a mixture while cutting off as much computations of other Gaussians as possible. These methods are described as follows:

k-best vector threshold Use the value of the temporal k-th best Gaussian as the pruning threshold. A Gaussian is pruned if the accumulated distance reaches the threshold while computing each distance component. If it is not pruned to the last dimension, it means that the value is within the k-best, so update the k-best threshold.

k-best vector threshold initiated by previous best Same as above but the k-best Gaussians of the previous frame are computed first. As input vectors change gradually in successive frames, we can expect that the best Gaussian set in the previous frame gets higher scores. This makes the initial threshold closer to the true k-best value.

vector threshold with heuristic estimation When computing a Gaussian, its expected value is estimated by adding the temporary maximum values of the yet-to-be-computed dimensions to the current accumulated distance. Pruning is performed by the estimated score. The initial maximum score on each dimension is set up by computing the previous k-best Gaussians first.

scalar threshold (dimension-independent pruning) An independent threshold is set up for each dimension by an offset from the maximum. First, compute the previous k-best Gaussians and get the maximum for each dimension. A Gaussian is pruned if its scalar value on the dimension is below a certain range.

The former two thresholds are *safe* in that the precise k-best Gaussians are guaranteed to be obtained without errors. The latter two are *unsafe*, rather aggressive methods where k-best Gaussians can be lost in the computation process by mis-leading of heuristics or using a too narrow score range.

4. EXPERIMENTAL RESULTS

The accuracy and efficiency of the proposed PTM model are evaluated. We build gender-independent PTM HMMs of 43 phones that have 3 states for each, and each state has a mixture of 64 Gaussians.

The task is 20k-word dictation of Japanese newspaper articles with a word trigram model. The acoustic model is integrated with JULIUS, our two-pass decoder based on A* search[3]. The reference models are gender-independent tied-state triphone models. They are all of 2000 states but different in number of mixture components per state. These modules are all available in Japanese dictation toolkit[4].

Test set contains 200 sentences spoken by 46 speakers, equal number of male and female, from Japanese speech corpus collected by Acoustical Society of Japan[5].

4.1. Comparison of Models

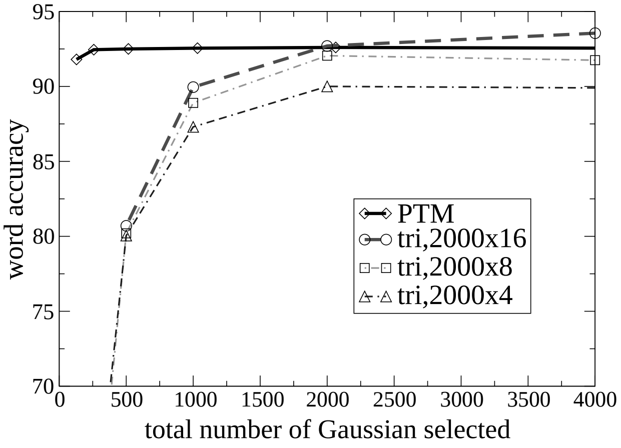
Word accuracy of the PTM model and triphone models is compared in Table 1. Scale factors of the models are also stated here. The proposed PTM achieves higher accuracy than the shared-state triphone of the same complexity (i.e. total number of Gaussian), and is comparable to the triphone of four times as many mixture components. The figure is almost best for the test set. Thus it is proved that the proposed PTM is superior to the triphone in that the same accuracy can be achieved with less parameters. The “PTM,synthesized” in Table 1 is a model that re-estimates only the mixture weights and does not re-train the mixture components. On the other hand, the “PTM,re-trained” model re-estimates both. The latter achieves better accuracy, thus it is shown that re-estimating not only mixture weights but also the shared Gaussian distributions is effective.

Table 1: Comparison of models

HMM model	state× mix. size	total G. #	accuracy
triphone	2000×16	32000	93.8
	2000× 8	16000	92.7
	2000× 4	8000	90.8
PTM,synthesized	129×64	8256	92.3
PTM,re-trained	129×64	8256	93.0

G. #: number of total mixture Gaussians
beam width = 1500

Next, we examine how these models are affected by Gaussian pruning. Accuracy of each model against various numbers of selected Gaussians is plotted in Figure 2. For comparison, Gaussians are pruned independently within each mixture in PTM and across all Gaussians in triphone models. A smaller beam width is used in decoding for convenience.



(beam width = 800)

Figure 2: Accuracy decrease by Gaussian selection

The proposed PTM keeps high accuracy even when the number of selected Gaussians is limited to only 3% (2-best out of 64 mixture components in a mixture, 258 in total). On the other hand, triphone systems are obviously sensitive to Gaussian pruning. Over 2000 Gaussians per frame were required to get sufficient accuracy. Our state-based PTM consists of distributions without redundant components, and makes severe pruning possible.

4.2. Comparison of Gaussian pruning methods

Next, we evaluate the performance of Gaussian pruning methods to reduce the total computation cost. In this comparison, we want to get the best two Gaussians out of 64 mixture Gaussians with cutting computation of others as much as possible. The computational cost is measured by the total percentage of computed Gaussian distance components.

Table 2: Comparison of Gaussian pruning methods

pruning methods	Gaussian distance components computed	acc.
k-best	59 %	92.5
k-best,previous-best	52 %	92.5
heuristic	36 %	92.3
scalar	21 %	92.2

beam width = 800, 2-best selected

In Table 2, both computed amount and word accuracy for the proposed methods are listed. The first vector-threshold method reduces the computed Gaussian distances to 59%. And by setting initial threshold by the previous k-best Gaussians, the ratio is improved to 52%. As pruning error never occurs in these methods, they are simple and reasonable ways to reduce the acoustic matching cost to a half without decreasing any accuracy.

Using the heuristic estimation reduces the cost further to 36% with little pruning error. The scalar threshold (dimension-independent pruning) realizes the best performance. The computed densities are remarkably reduced to 21%, with little loss of accuracy. But as the pruning performance is determined by the threshold range, the scalar method needs sensitive tuning.

5. MONOPHONE LEXICON TREE ON PRELIMINARY RECOGNITION

As the proposed PTM HMMs are built using mixtures of monophone HMMs, it is possible to re-define monophone models together with triphones by sharing the same mixture components. As monophones are context-independent, we can make a smaller lexical tree and omit handling context dependency. So we explore the possibility of using a monophone tree lexicon at the preliminary recognition in our multi-pass decoder for further efficiency.

The result is shown in Table 3. The compaction of the lexicon tree does not improve the speed, and the growing errors on the preliminary pass decrease the accuracy on the final result. The results confirmed that the use of better acoustic model on the first pass is significant.

Table 3: Lexicon tree: triphone vs. monophone

lexicon	state #	acc. (acc1)	time(×RT)
triphone	173251	92.2 (82.2)	4.5
monophone	128188	90.2 (76.8)	4.4

acc1: accuracy on the preliminary pass
CPU: UltraSPARC 300MHz

Finally, by tuning search parameters and using a smaller beam width, accuracy of 90.4% is achieved with a speed of 2.3 times the real time.

6. CONCLUSION

A new PTM model with state-based mixture-tying scheme has been introduced. The model is synthesized from mixtures of a monophone model and state-clusters of triphone models. As the construction of the model is straightforward, training of the parameters can be more reliable.

This model achieves a word error rate of 7.0%, which is comparable to the best figure by the triphone of much higher complexity. With Gaussian pruning, computing only 2 out of 64 mixtures per state does not cause any loss of accuracy. The acoustic computational cost is reduced to about 20% by the dimension-independent pruning.

Acknowledgment: The work is sponsored by IPA (Information-technology Promotion Agency), Japan. The authors are grateful to members of the project for their comments and cooperation.

7. REFERENCES

[1] G.Zavaliagkos, J.McDonough, D.Miller, A.El-Jaroudi, J.Billa, F.Richardson, K.Ma, M.Siu, H.Gish: The BBN BYBLOS 1997 Large Vocabulary Conversational Speech Recognition System, In *IEEE ICASSP*, pages 905-908, 1998.

[2] A.Sankar: A New Look at HMM Parameter Tying for Large Vocabulary Speech Recognition, In *Proc. ICSLP*, pages 2219–2222, 1998.

[3] A.Lee, T.Kawahara, S.Doshita: An Efficient Two-Pass Search Algorithm using Word Trellis Index, In *Proc. ICSLP*, pages 1831–1834, 1998.

[4] T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, M.Yamamoto, A.Tamada, T.Utsuro and K.Shikano: Sharable Software Repository for Japanese Large Vocabulary Continuous Speech Recognition, In *Proc.ICSLP*, 1998.

[5] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, K.Shikano and S.Itahashi: The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus, In *Proc.ICSLP* 1998.