GAUSSIAN MIXTURE SELECTION USING CONTEXT-INDEPENDENT HMM

Akinobu Lee * *Tatsuya Kawahara* † *Kiyohiro Shikano* *

* Nara Institute of Science and Technology, Ikoma 630-0101, Japan
† Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

We address a method to efficiently select Gaussian mixtures for fast acoustic likelihood computation. It makes use of context-independent models for selection and back-off of corresponding triphone models. Specifically, for the k-best phone models by the preliminary evaluation, triphone models of higher resolution are applied, and others are assigned likelihoods with the monophone models. This selection scheme assigns more reliable back-off likelihoods to the un-selected states than the conventional Gaussian selection based on a VQ codebook. It can also incorporate efficient Gaussian pruning at the preliminary evaluation, which offsets the increased size of the pre-selection model. Experimental results show that the proposed method achieves comparable performance as the standard Gaussian selection, and performs much better under aggressive pruning condition. Together with the phonetic tied-mixture (PTM) modeling, acoustic matching cost is reduced to almost 14% with little loss of accuracy.

1. INTRODUCTION

In recent studies of large vocabulary continuous speech recognition, large-scale HMM containing a large number of Gaussians has been adopted as acoustic modeling to achieve accurate recognition. However, computing a huge number of Gaussians increases the acoustic matching cost enormously. In most recognition systems, the acoustic computation cost often occupies most of the decoding time. Thus, an efficient method to compute the acoustic likelihood with such a large-scale model has been a major concern for practical real-time recognition systems.

Gaussian Selection (GS) is one approach widely adopted in various large vocabulary continuous speech recognition systems. A VQ codebook is trained and all the defined Gaussians are clustered according to the codebook beforehand. In decoding, instead of calculating all the Gaussians needed, only the Gaussians within the cluster nearest to the input vector are computed. As it can remarkably reduce the amount of acoustic computation, many extensions and variations have been studied so far[1][2][3][4].

However, this kind of pruning approach for fast decoding has an essential problem that not a few states, whose mixture components are all pruned, have entirely no value. The performance largely depends on the pruning threshold (i.e. cluster size in GS). Under aggressive condition with a tight threshold, the accuracy decreases remarkably. Assigning some constant value to those pruned states eases this error, but this flooring method is not the best solution. This problem is inevitable for all GS schemes.

In this paper, we propose a Gaussian mixture selection method based on likelihood of context-independent HMM. Instead of training a VQ codebook, we uses state probabilities of a context-independent HMM to select corresponding triphone states, i.e., Gaussian mixtures. All the context-independent HMM states are computed first, and only the triphone states whose corresponding monophone states are ranked within the k-best are computed. The unselected states are given the probability of monophone itself. This approach is advantageous in that 1) the unselected states are reliably "backed-off" by assigning actual likelihood of corresponding monophone probabilities, and that 2) selecting Gaussians per a mixture enables Gaussian pruning that can further reduce the computational cost. These features realize stable recognition with even more tight condition.

2. GAUSSIAN SELECTION

Gaussian Selection (GS) is a popular approach for fast likelihood calculation. When an input vector lies on the tail of a Gaussian distribution, the output probability of the Gaussian is very small. It results in a tendency that only several Gaussians nearest to the input have dominant effect on the final output probability of a state, and ignoring those far from the input vector will not affect the recognition accuracy. So instead of evaluating all the Gaussians, computing only the Gaussians near the input vector will be sufficient. The GS methods try to select such Gaussians efficiently for an input vector.

A standard GS method is based on vector quantization (VQ), originally proposed by Bocchieri[1]. The acoustic space is divided into a set of vector quantized regions, and all Gaussians are clustered to one or more VQ codewords. When recognition, the input vector is quantized to a single VQ codeword and only the Gaussians within the cluster assigned to the codeword are computed. The Gaussians not in the cluster are "pruned" and not calculated at all. To avoid mis-selection and to control the amount of selection, the Gaussians are shared among clusters according to the distance from codewords. A Gaussian belongs to a cluster if its Euclidean distance from the codeword vector is below a given threshold. In this paper we use variance-weighed distance function for clustering as a baseline[3]. The reduction of computational cost is dependent on the size of the cluster, and there is a trade-off between the size and recognition errors.

One of vital problems for the GS scheme is a state flooring. The clusters are defined based on acoustic space partitioning independently from the state and mixture structures. As only Gaussians in the selected cluster are computed, not a few states that have no Gaussian components included in the selected cluster gets absolutely no value. As a result, associated hypotheses are forced to be removed from decoding. To avoid pruning errors, these states have to be "floored" with a kind of approximate value. A simple solution for the state flooring is to assign a constant value to those states. However, the discrete flooring does not reflect actual likelihood of the input, and assigning such unreliable values is crucial to maintain high accuracy. Furthermore, under more aggressive condition with smaller clusters, many states are floored and the accuracy decreases remarkably.

To deal with such floored states is an essential problem in all other GS methods based on acoustic partitioning. Although several enhancements on clustering have been proposed such as limiting maximum number of Gaussians per state[3], they do not solve the problem fundamentally.

3. MIXTURE SELECTION USING CONTEXT-INDEPENDENT HMM

Instead of making Gaussian clusters, we propose a per-state mixture selection method using a simple context-independent HMM and its hierarchical correspondence with triphone models. Figure 1 illustrates the overview of the procedure. We assume that all triphone models of the same center phone have the same number of states as corresponding monophone model, and that they are trained with the same training data. All monophone states are evaluated first for every input frame, and the k-best states are determined.

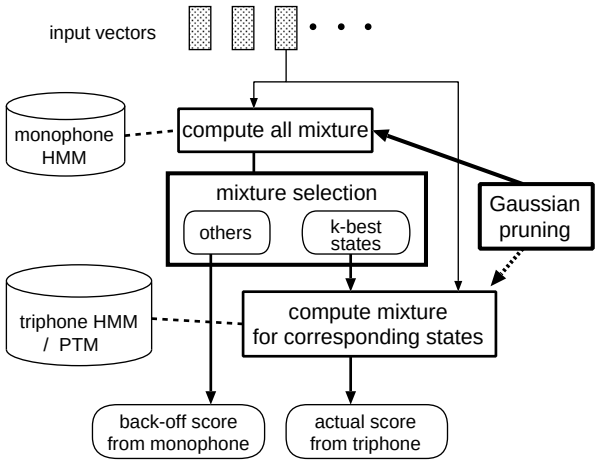

Fig. 1. Mixture Selection using Context-Independent HMM

Then, only Gaussian mixtures of triphone states that correspond to the k-best monophone states are computed. Scores of triphone states whose corresponding monophone states get lower than the k-best, are "backed-off" by assigning the likelihood of the monophone model itself.

We call this method Gaussian mixture selection. Since monophone models are trained with the maximum likelihood criterion, they serve as good approximation of triphones than a mere constant value or other ad-hoc computed values. Their scores reflecting the actual input makes back-off scoring of un-selected states work even with aggressive pruning with few selected states.

Another advantage of the proposed method is its easy and straight-forward application. Training and implementation of the preliminary selection scheme is very simple. Only the monophone model trained with the same corpus is needed, and as it is often generated in the way of building a triphone model, no extra training is actually needed, in contrast with the conventional GS that needs training of an optimum VQ codebook and clustering a vast number of Gaussians. The proposed selection method can also be implemented with a little modification on the recognition system.

However, the scheme has not been widely adopted due to the increased computation of cost of pre-selection itself. Although it is desirable to use large monophone models with sufficient Gaussians for more accurate selection and back-off, the increase of matching cost for preliminary evaluation results in much more computational cost, sometimes spoils the selection effect itself.

To solve this trade-off problem, we further incorporate another selection method called Gaussian pruning[6]. Given an input vector and a Gaussian set, it computes

only k-best Gaussians while dynamically dropping off unpromising ones during accumulation of distances for each vector element. We have presented and compared several implementation methods for the phonetic tied-mixture model[6]. The pruning algorithm drastically reduces the computation of monophone models for pre-selection with little loss of selection accuracy, which does not affect the final result. Moreover, it is also applicable to evaluation of selected triphone models, as the selected states have full Gaussian mixtures. Notice that introducing further pruning in the cluster of conventional GS is not admissible and only worsen the state flooring problem.

4. EXPERIMENTAL RESULT

The proposed Gaussian mixture selection is evaluated in comparison with the standard Gaussian selection. We implemented both methods on recognition engine Julius[5], our two-pass decoder based on A* search.

The task is 20k-word recognition of Japanese newspaper article corpus with a word trigram model. Two kinds of gender-dependent acoustic models are prepared for evaluation. One is a tied-state triphone model of 2000 states, in which each state has a mixture of 16 Gaussians. The other is a phonetic tied-mixture (PTM) model [6] in which mixtures are shared among triphone states of the same position of the same base phone. In total, 129 codebooks are defined for the PTM and each has 64 Gaussians, and they are shared with different weights among 3000 states. Test set contains 100 sentences spoken by 23 female speakers. These modules are all available in Japanese dictation toolkit[7]. For SGS, a codebook of 1119 Gaussians is set up.

First, we compare the proposed Gaussian mixture selection (GMS) with conventional standard Gaussian selection (SGS) with the tied-state triphone model. To control the number of Gaussians to be selected, we set up several sets of clusters for SGS, which are different in sizes and controlled by the distance thresholds. In GMS, monophone model that has 16 mixtures in each state is used and Gaussian pruning is applied on the selection process to get only the maximum Gaussian probability. The selection parameter in GMS is controlled on run time by specifying the number of monophone states to be selected.

The average number of computed Gaussians per frame and the accuracy at different selection sizes are listed in Table 1. Only the Gaussians actually computed are counted in triphone. For fair comparison, preliminary computation in the selection procedure is also counted. In SGS, computational cost of calculating distances to the cluster centroids (codewords) is added in terms of the number of Gaussian likelihood computation. On GMS, the calculation of likelihoods of monophone HMM for preliminary selection is added.

Table 1. Comparison of methods with tied-state triphone

GS method		#Gauss. tri	#Gauss. pre	total %Gauss.	word %Err.
no GS		15772	—	100.00	4.5
SGS	2.1	6672	1119	49.40	4.5
	1.7	4132	1119	33.29	5.2
	1.3	2222	1119	21.18	6.2
	0.9	971	1119	13.25	15.7
GMS	48	6660	690	46.60	5.1
	24	3712	690	27.91	5.9
	8	1468	690	13.68	6.4
	4	824	690	9.60	8.6

SGS parameter: distance threshold for clustering
GMS parameter: number of monophone states to be selected out of total 129 states
tri: computed Gaussians in triphone
pre: cost in preliminary selection

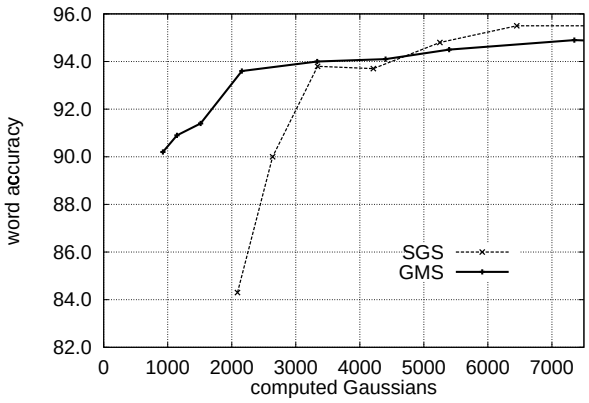

Fig. 2. Selection performance of SGS and GMS

The proposed GMS method achieves comparable performance to the conventional SGS given a sufficient number of computed Gaussians. Furthermore, it works more efficiently and stably at a small number of selected mixtures. The accuracy against the number of computed Gaussians is plotted in Figure 2. Since the back-off scores for un-selected states are more reliable than SGS, it is confirmed that GMS does not lose accuracy so much even with a tight threshold.

The computation overhead by selection in GMS is smaller than SGS. Although the monophone HMM has 2064 Gaussians in total, introducing Gaussian pruning in the preliminary selection reduced the computational cost to 690.

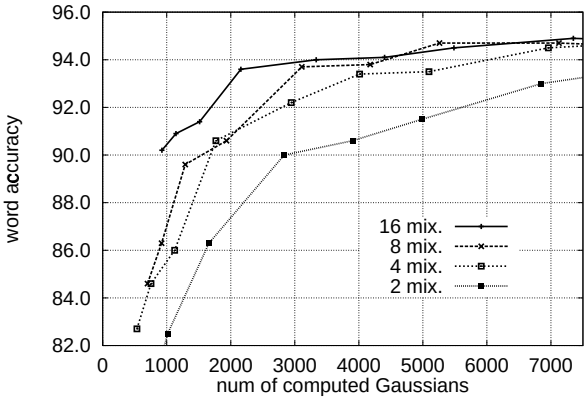

Fig. 3. Comparison of selection models

Table 2. Effect of GMS on PTM model

GS method	#Gauss.		total	word
	tri	pre	%Gauss.	%Err.
triphone,2000x16	2644	690	21.14	5.9
PTM,129x64	434	690	13.61	6.0

selection model: 16mix. monophone

Next, we compare several selection models of different sizes. The costs and accuracies against various number of selected states are plotted in Figure 3. As a smaller model with fewer Gaussians has less back-off ability, the recognition accuracy decreases to a large extent. This result confirms that assigning good back-off likelihoods to the unselected states is significant. The selection cost is higher in larger monophones, but the Gaussian pruning on the preliminary selection offsets the increase of model size.

The performance of the proposed GMS together with the phonetic tied-mixture (PTM) model is shown in Table 2. Even with the PTM model, whose parameter size is already small, the number of computed Gaussians was reduced to 13.61% with little accuracy decrease. The PTM model combined with GMS achieves comparable accuracy to the standard triphone model with only 1124 Gaussians per frame computed.

Finally, we seek for the best performance of the system by tuning search parameters and using a smaller beam width. Test set is now extended to 200 sentences by 46 speakers, equal number of male and female. As a result, the error rate of 7.8% is achieved in real time decoding with a standard PC.

5. CONCLUSION

An efficient method to select Gaussian mixtures for fast likelihood calculation is presented. The state likelihoods of context-independent model are used for both state selection and back-off. It gives reliable scoring of pruned (un-selected) triphone states, thus realizes more efficient recognition under the aggressive pruning condition than the conventional VQ-based Gaussian selection. The property will be advantageous for robust recognition in mis-matched condition. Together with the phonetic tied-mixture (PTM) modeling, acoustic matching cost is reduced to almost 14% with little loss of accuracy.

Acknowledgment: Part of the work is sponsored by CREST (Core Research for Evolutional Science and Technology), Japan.

6. REFERENCES

[1] E.Bocchieri: Vector Quantization for Efficient Computation of Continuous Density Likelihoods, In *Proc. IEEE-ICASSP*, pages 692–695, 1993.

[2] K.M.Knill, M.J.F.Gales and S.Young: Use of Gaussian Selection in Large Vocabulary Continuous Speech Recognition using HMM's, In *Proc. ICSLP*, pages Vol.1, pages 470–473, 1996.

[3] M.J.F.Gales, K.M.Knill and S.J.Young: State-Based Gaussian Selection in Large Vocabulary Continuous Speech Recognition using HMM's, In *IEEE Trans. on Speech and Audio Processing*, Vol.7, No.2, pages 152–161, 1999.

[4] D.B.Paul: An Investigation of Gaussian Shortlists, Proc. IEEE workshop on Automatic Speech Recognition and Understanding, 1999.

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

[6] A.Lee, T.Kawahara, K.Takeda and K.Shikano: A New Phonetic Tied-Mixture Model for Efficient Decoding, In *Proc. IEEE-ICASSP*, pages 1269–1272, 2000.

[7] T.Kawahara, A.Lee, T.Kobayashi, K.Takeda, N.Minematsu, S.Sagayama, K.Itou, A.Ito, M.Yamamoto, A.Yamada, T.Utsuro and K.Shikano: Free Software Toolkit for Japanese Large Large Vocabulary Continuous Speech Recognition, In *Proc. ICSLP*, Vol.4, pages 476–479, 2000.