DISCRIMINATIVE DYNAMIC GAUSSIAN MIXTURE SELECTION WITH ENHANCED ROBUSTNESS AND PERFORMANCE FOR MULTI-ACCENT SPEECH RECOGNITION

Chao Zhang[⋆]*, Yi Liu*[⋆]*, Yunqing Xia*[⋆]*, Chin-Hui Lee*[†]

[⋆]Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, China
[†]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

We propose a discriminative DGMS (dynamic Gaussian mixture selection) strategy to enhance restructuring of a pre-trained set of Gaussian mixture models to cover unexpected acoustic variations at run time in automatic speech recognition. The number of Gaussian components in each hidden Markov model (HMM) state set aside is determined by a minimum classification error criterion. We also use a genetic algorithm to solve the integer programing problem to find the globally optimal state size. This parameter is used to adjust the HMM state densities for each input speech frame, leading to both high robustness and good resolution for dynamic tracking to cover a diversity of temporal variations in speech. Tested on an accented speech recognition application, the proposed framework yields an improved syllable error rate reduction over the conventional DGMS and augmented HMM systems when evaluated on three typical Chinese accents, Chuan, Yue and Wu, while maintaining its performance for standard Putonghua.

Index Terms— Dynamic Gaussian Mixture Selection, Minimum Classification Error, Genetic Algorithm, Accent

1. INTRODUCTION

Conventional dynamic Gaussian mixture selection (DGMS) [1] empirically determines a parameter vector, with each integer element indicating the size of a Gaussian mixture model (GMM) used to characterize one particular state from a subset of hidden Markov model (HMM) states set aside for covering unknown acoustic variations in speech recognition for dynamic frame-level GMM reconfiguration at run time. In this paper we propose a discriminative DGMS strategy to determine the parameters with a minimum classification error (MCE) [2] criterion over a training set containing condition-specific data able to cover some intended variations. We also propose solving this integer programming problem with a genetic algorithm (GA) [3] to find a globally optimal combination of GMM sizes by a heuristic to avoid going over its entire search space, and thus greatly reducing the cost.

We also test the proposed DGMS strategy on accented speech recognition in which the multi-accent variations can be accounted for in the conventional DGMS algorithm [1]. Accent is a crucial bottleneck for an extensive usage of speech-enabled applications across a large population in China since all Chinese speakers share the same ideographic characters but with different pronunciations due to regional accents. There are seven major dialectal regions in China: Guanhua, Yue, Wu, Xiang, Gan, Min, and Kejia, which have quite different acoustic and linguistic representations from Putonghua [4]. Hence, speakers whose first language is a native dialect have their Putonghua pronunciations inevitably influenced by the dialect. Since accented speech differs from the standard one in terms of acoustical and phonological characteristics, most state-of-the-art Putonghua ASR systems fail to perform well when the speaker has a regional accent. This problem is especially severe when multiple accents are presented.

Traditional methods for accented speech recognition focus on handling accent variations at the acoustic and phonetic levels [4-5]. Besides the straightforward method of building models with a large amount of accented speech, maximum a posteriori (MAP) [4] is commonly used to adapt the standard models to fit certain accent to cover potential acoustic changes [4-5]. A major weakness of conventional adaptation approaches [5] is that the parameters of the acoustic models undergo an irreversible change, making the models losing their ability to cover other accents and standard speech. An approach to state-level pronunciation modeling with model reconstruction was thus proposed to handle both acoustic and phonetic changes for accented speech [4]. However, the method without optimization increases the model size, resulting in inefficient use of the Gaussian components and causing model resolution degradation. This inevitably brings serious performance degradation to beam pruning in the ASR decoding process.

Targets at multi-accent speech recognition [1], we propose discriminative DGMS with each reconstructed HMM state having an individual number of Gaussian components that are the nearest and most representative to the current frame. Experimental results show that our proposed DGMS method is robust in pruned beam search and yields a significant absolute syllable error rate reduction by 11.17%, 11.96% and 10.88%, respectively, on the *Chuan*, *Yue*, and *Wu* accents from the conventional *Putonghua* triphone HMM system, while maintaining its performance on standard Putonghua.

2. ACOUSTIC MODEL RECONSTRUCTION FOR HMM AUGUMENTATION

Acoustic model reconstruction at the state level was proved capable of handling accent changes at both acoustic and phonetic levels in our previous work [4]. In general, the procedure of traditional model reconstruction is achieved with the following five steps on the accented data:

1) **Obtain a canonical transcription with time**. This is achieved by forced alignment on the phone-level canonical transcriptions.

2) **Obtain an alternative transcription** by phone classification according to the time boundaries obtained in step 1) [1].

3) **Extract the reliable accent specific unit** (ASU) by comparing the canonical and alternative transcriptions. Each reliable ASU is selected from the misclassified phones and represents an accent

change [1].

4) **Estimate an auxiliary tree for reliable ASU** by creating its tied-state triphone models by decision tree-based clustering with its relevant instances selected from the misclassified phones [1].

5) **Reconstruct the canonical acoustic models to cover accent changes** [4]. Accented Gaussian components from a node of auxiliary decision tree are borrowed to enlarge and adjust the Gaussian mixture distribution of a leaf node on its relevant canonical decision tree (illustrated in Figure 1) [4].

The above approach seeks to cover accent changes by adjusting canonical Gaussian mixture densities with the borrowed accented Gaussian components. We refer to the states adjusted by acoustic model reconstruction as the statically reconstructed states to distinguish themselves from the ones dynamically reconstructed by DGMS. More details about reliable ASUs and acoustic model reconstruction are given in our previous studies [1, 4].

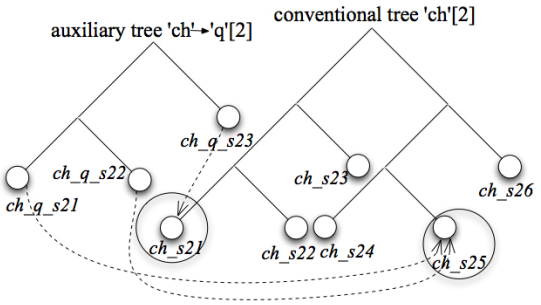

Fig. 1. Decision tree merging for ASU 'ch'→'q', where 'ch' is the canonical phone, and 'q' an alternative phone.

3. MCE BASED DISCRIMINATIVE DYNAMIC GAUSSIAN MIXTURE SELECTION

Acoustic model reconstruction brings advantages with a price of increasing the acoustic model size. Meanwhile, it can be seen from Figure 2 (A) and (B) that integration of the Gaussian components may degrade model resolution. For an extreme case, the degradation is serious in beam search pruning during decoding because all the components are considered in likelihood calculation.

To overcome the model resolution degradation without increasing model size, for each decoded speech frame, we use DGMS to dynamically reconstruct the output density of a statically reconstructed state to compute its observation likelihood. The parameters that control DGMS to reconstruct the output densities are pre-determined discrete variables, which are optimized by GA according to MCE criterion. Therefore, we obtain discriminative DGMS.

3.1. Dynamic Gaussian Mixture Selection (DGMS)

To avoid model resolution loss and to increase robustness on accent variations, in decoding, DGMS adjusts the output density of a statically reconstructed state by selecting k components being nearest to an input frame. A dynamic output distribution is thus built to compute the observation likelihood for that frame during decoding. DGMS is explained formally as follows.

Let $N_m = \mathbf{N}(\mu_m; \mathbf{\Sigma}_m)$ denote the m-th Gaussian mixture, $b_r(\mathbf{o}) = \sum_{m=1}^{M} w_m N(\mathbf{o}|\mu_m; \mathbf{\Sigma}_m)$ the output density for the reconstructed state, r, $N_1^{'}, \ldots, N_k^{'}$ the k Gaussian components nearest to speech frame $\mathbf{o}$ in terms of Mahalanobis distance, and the dynamical output distribution is evaluated as follows.

$$\begin{cases} b_r^{'}(\mathbf{o}) = \sum_{m=1}^{k} w_m^{''} N(\mathbf{o}|\mu_m^{'}; \mathbf{\Sigma}_m^{'}) \\ w_m^{''} = w_m^{'} \Big/ \sum_{m=1}^{k} w_m^{''} \end{cases} . \quad (1)$$

For the acoustic samples of an accented speech frame that are located at the boundary of the Gaussian densities, DGMS chooses k Gaussian mixtures being most representative to the relevant accent changes, and the obtained dynamical output density has a better model representation ability as illustrated in Figure 2 (C). Meanwhile, for acoustic samples of the standard speech frames located at the center of the density, its dynamical output density is similar to that without model reconstruction, as shown in Figure 2 (D), and retains its covering ability for standard speech.

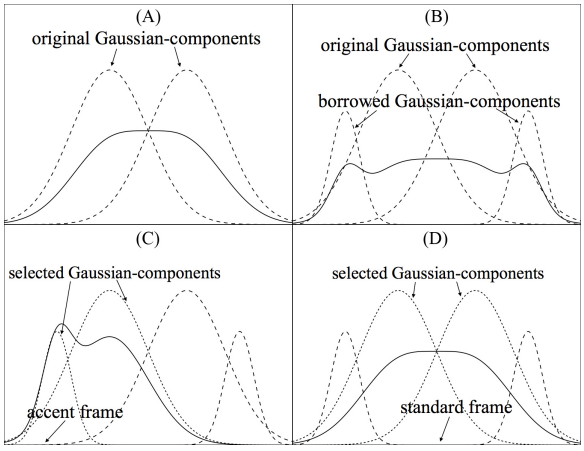

Fig. 2. Output densities of model reconstruction and DGMS.

To provide sound flexibility to fit the diversified accent changes, k is pre-specified and differentiated for different HMM states. However, this gives many statically reconstructed state parameters to DGMS. We refer to them as the parameter vector. Difficulties in finding the optimal parameter vector are: 1) how to evaluate a vector; 2) how to find the optimum efficiently.

3.2. MCE Criterion for Discriminative Dynamic Gaussian Mixture Selection

MCE is a widely used criterion to minimize an estimation of the errors in the training set. Its effectiveness has been shown in discriminative training of HMMs for ASR [2, 7-8]. We adopt MCE to select the parameter vectors and obtain discriminative DGMS.

Suppose O is the feature sequence of a training utterance, and Λ is the set of parameters for acoustic models, MCE is formulated with discriminant functions $g_j(O, \Lambda, c)$ [2, 6] which represent the acoustic log-likelihood for an output string S_j scored with DGMS using a parameter vector c. A misclassification measure $d_i(O, \Lambda, c)$ is used to evaluate the acoustic difference between the canonical transcription S_i and its alternative transcriptions.

$$d_i(O, \Lambda, c) = -g_j(O, \Lambda, c) + \max_j g_j(O, \Lambda, c). \quad (2)$$

Therefore, the loss function for MCE can be defined as follows.

$$l(d_i(O, \Lambda, c)) = \frac{1}{1 + e^{-\alpha \cdot d_i(O, \Lambda, c)}}. \quad (3)$$

We use the MCE loss function as the objective function in optimizing the parameters of DGMS. When the minimum value in Eq. (3) is obtained, DGMS minimizes the empirical risk by reducing a maximum number of training set errors, including those caused by accent. Therefore, it increases the covering ability for accent variations. Moreover, by assigning different weights to training utterances with different correct/incorrect degrees we are able to control the optimization using varying values of α [6].

The canonical and alternative transcriptions are generated by forced alignment in free grammar recognition, keeping track of the full state alignment. We use the phone sequences that score the largest among fixed n-best outputs, to approximate the exact alternative transcriptions for vector c.

3.3. Solving Optimal Parameter Vector by Genetic Algorithm

To obtain the optimal value of Eq. (3) by varying parameter vectors, the entire search space grows exponentially along with the vector size increases. It is often not feasible to find the optimum by enumerating all possible vectors. Therefore, we use GA to perform the optimization process efficiently to solve for the vector c.

GA is a random "search for solution" algorithm that is capable of finding the optimal solution by examining over only a small fraction of the possible candidates via mimicking the survival of the fittest process of natural evolution [3]. Viewing a DGMS parameter vector as a chromosome, tuning for optimum parameters equals to find the fittest chromosome during the evolution. Therefore, a chromosome c is constituted by R positive integers being relevant to the Gaussian component selection number k for the R statically reconstructed states. Every integer ranges from a pre-specified minimum to the Gaussian component number in its relevant state that is reconstructed statically.

We use MCE to define the fittest function, $f(c)$, of GA. Note $f(O,c) = l(d_i(O,\Lambda,c))$, $f(c)$ is the average of $f(O,c)$ over the entire training set. GA for DGMS works as follows:

1) Generate N chromosomes randomly as the initial population;

2) Compute a fitness function $f(c)$, for every chromosome;

3) Select C chromosomes randomly from the current population, C is even. The selection probability is

$$P(c) = \frac{1}{f(c)} \Big/ \left(\sum_{c'} \frac{1}{f(c')} \right), \qquad (4)$$

where $c^{'}$ refers to the set of chromosomes in the population. We use a conventional "roulette-wheel sampling" method to make the selection as explained in details in [3].

4) Reproduce the selected chromosomes by sequentially dividing them into $C/2$ pairs. "One-point crossover" is used for reproduction of each pair [3], and generates C children.

5) Merge the child chromosomes generated from step 4) with the unselected ones from step 3), and form a new population. Generate a random probability for each chromosome at each locus, if the random probability is smaller than a pre-set mutation probability, the integer at the current locus is randomly increased or decreased by one with equal probability. Replace the original chromosomes with their variations.

6) Replace the current population by the new population.

7) Repeat steps 2-6, if no c satisfies $f(c) \approx 0$ and the maximum generation number is not reached.

The chromosome with the smallest fittest function value is the optimum parameter vector for DGMS.

Remarkably, we use GA instead of other "search for solution" algorithms for the following two reasons: 1) Chromosomes of GA have inherently discrete representations, which match our goal of optimizing discrete variables; 2) GA is guaranteed to reach the global optimum if enough generations have been reproduced [3].

4. RECOGNITION EXPERIMENTS

The 863 regional accent speech corpus [7] was used in our experiments to evaluate our method on three typical accents – Chuan, Yue, and Wu. This database is the largest one and most commonly used in Chinese accented speech recognition [1]. All data were sampled at 16kHz with a 16-bit precision. There is no word n-gram in these sentences so that we can isolate the effect of our method without the influence from high-level information. Table 1 lists the detailed statistics for the datasets.

Table 1. Data set separation in all experiments

Data ID	Duration	Syllable Number	Speaker Number	Utter. Number	Type
DevC	6.5h	51,907	20	3,205	Chuan
TestC	4.3h	33,847	20	2,000	
DevY	6.1h	51,341	20	3,091	Yue
TestY	3.5h	31,191	20	2,000	
DevW	6.6h	52,584	20	3,471	Wu
TestW	3.8h	29,888	20	2,000	
TrainP	51.5h	340,556	100	25,920	Putonghua
TestP	3.9h	23,158	10	2,000	

The HMM topology was three-state, left-to-right without skips. The acoustic features were 39-dim vector with $13MFCC$, $13\Delta MFCC$, and $13\Delta\Delta MFCC$. 28 initials and 36 finals were used to generate context-independent HMMs. Our Putonghua triphone HMMs with 3,000 tied-states and 12 Gaussian components per state were trained using Putonghua data set TrainP by HTK decision tree based state-tying. A dictionary with all 413 syllables was used.

165, 191, and 116 reliable ASUs were extracted from DevC, DevY, and DevW, respectively, resulting in 495, 573, and 498 auxiliary trees with 517, 605, and 533 tied-states, respectively. The augmented acoustic models include 42,620 Gaussian components (14.2 components per state on average). Our augmented HMMs have 546 statically reconstructed states. α for MCE was set to 0.01 to map many training tokens falling into the linear region of the loss function in Eq. (3), and to increase the separation between correct/incorrect hypotheses as well as the generalization ability on the unseen accented data [6]. For GA in our experiments, the population size N, reproduction size C, and mutation probability were set to 400, 200 and 0.3, respectively. The 5-best outputs obtained by the Augmented HMMs were used to approximate the exact alternative transcriptions in MCE. The optimization was forced to stop at a maximum of 1,200 iterations, and its objective function was converging and reduced from 0.86 to 0.62.

We will first examine the effectiveness in using DGMS to avoid model resolution loss. We evaluate model resolution by

$$D_i(\Lambda) = -\sum_{O} d_i(O,\Lambda,c)/g_i(O,\Lambda,c), \qquad (5)$$

where g_i is the acoustic log-likelihood with on canonical state i; d_i is the same as in Eq. (2) that represents the acoustic difference for a observation sequence scored on its canonical and alternative states.

Therefore, the model resolution for state i increases when $D_i(\Lambda)$ is increased. Let $D_i(\Lambda_0)$ denote the resolution of the augmented HMMs, $D_i(\Lambda^*)$ the resolution with DGMS. Hence, a relative model resolution improvement obtained by DGMS is

$$\text{sgn}(D_i(\Lambda^*)) \cdot |D_i(\Lambda^*)/D_i(\Lambda)_0|. \quad (6)$$

Relative model resolution improvements for some selective tied-states are illustrated in Figure 3, which shows DGMS significantly improved the model resolution in Eq. (6), i.e., enhancing the separation between the canonical states and their alternative states. Therefore, in beam search, when the pruning threshold t, the value such that models whose maximum acoustic scores fall below it are deactivated, decreases, the search paths reserved by DGMS are more accurate than those reserved without DGMS, relieving the performance degradation in pruned search caused by acoustic model reconstruction, as we have presented in Figure 4. From Figure 4 we can see, with DGMS, system performance degraded slower as well as obtained lower SER when t decreases.

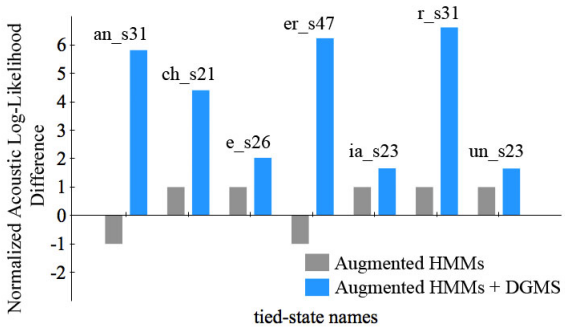

Fig. 3. Relative model resolution for representative states with/without DGMS evaluated on Yue accent, $D_i(\Lambda_0)$ is normalized.

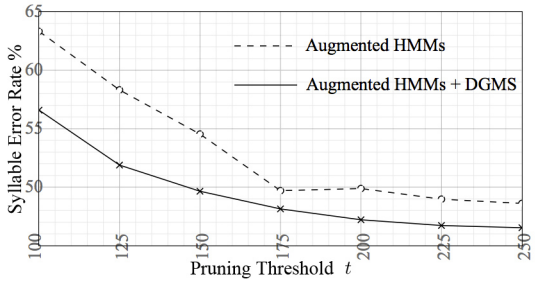

Fig. 4. SER variations for Augmented HMMs with/without DGMS, when pruning threshold t changes from 100.0 to 250.0 evaluated on Yue accent.

Finally, we will examine the recognition accuracy with the use of discriminative DGMS on multi-accent. As seen from Table 2, accents degrade recognition accuracy when the acoustic models were trained by merely standard-accent speech. The augmented HMMs relieve the performance degradation on multi-accent by borrowing accented Gaussian components to adjust pre-trained mixture distributions and enables more Gaussian components located at the mixture boundaries to cover accent changes [4], with little impact on Putonghua. When using DGMS with the augmented HMMs, SERs on Chuan, Yue, and Wu dropped relatively by 3.63%, 4.04%, and 3.94%. This improvement was achieved because minimizing classification errors increases the coverage abilities for accent changes. Meanwhile, our method maintained the performance on Putonghua as we restricted the reconstructed output densities with no less than 6 Gaussian components by DGMS, leading to retaining the pre-trained mixture distributions for standard speech.

Table 2. Lower SER for using discriminative DGMS to conventional triphone and Augmented HMMs

System	Syllable Error Rate (SER) %			
	TestP	TestC	TestY	Test W
Putonghua HMMs	22.48	56.57	58.31	55.46
Augmented HMMs	22.11 (-0.37)	47.11 (-9.46)	48.30 (-10.01)	46.41 (-9.05)
Augmented HMMs +DGMS	22.18 (-0.30)	45.40 (-11.17)	46.35 (-11.96)	44.58 (-10.88)

5. SUMMARY

Unexpected acoustic variations occur constantly at run time in automatic speech recognition. In this paper, a discriminative dynamic Gaussian mixture selection strategy is proposed to enhance restructuring of a pre-trained set of Gaussian mixture models so as to cover the variations. Experimental results show the new strategy is effective. Meanwhile, we conducted experiments on an accented speech recognition application, and results indicate the proposed framework yields an improved syllable error rate reduction over the conventional DGMS and augmented HMM systems when evaluated on three typical Chinese accents, Chuan, Yue and Wu, while maintaining its performance for standard Putonghua.

6. ACKNOWLEDGEMENT

This work was supported by Natural Science Foundation of China (60975018), State Scholarship Fund of China Scholarship Council (2010811231), and Nokia-Tsinghua Joint Funding 2008-2010.

7. REFERENCES

[1] C. Zhang *et al*, "Reliable accent specific unit generation with dynamic Gaussian mixture selection for multi-accent speech recognition," in *Proc. ICME*, pp. 1-6, 2011.

[2] B.-H. Juang, C. Wu, C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Process.*, vol. 5, pp. 257-265, May 1997.

[3] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA, 1999.

[4] Y. Liu and P. Fung, "Partial change accent models for accented Mandarin speech recognition," in *Proc. ASRU*, p. 111-116, 2003.

[5] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Process.*, vol. 2, pp. 291-298, Apr. 1994.

[6] E. McDermott and T. J. Hazen, "Minimum classification error training of landmark models for real-time continuous speech recognition," in *Proc. ICASSP*, pp. 937-940, 2004.

[7] A. Li *et al*, "RASC863-A Chinese speech corpus with four regional accents," in *Proc. Oriental-COCOSDA*, 2004.