Reliable Accent-Specific Unit Generation With Discriminative Dynamic Gaussian Mixture Selection for Multi-Accent Chinese Speech Recognition

Chao Zhang, *Student Member, IEEE*, Yi Liu, *Member, IEEE*, Yunqing Xia, *Member, IEEE*, Xuan Wang, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—In this paper, we propose a discriminative dynamic Gaussian mixture selection (DGMS) strategy to generate reliable accent-specific units (ASUs) for multi-accent speech recognition. Time-aligned phone recognition is used to generate the ASUs that model accent variations explicitly and accurately. DGMS reconstructs and adjusts a pre-trained set of hidden Markov model (HMM) state densities to build dynamic observation densities for each input speech frame. A discriminative minimum classification error criterion is adopted to optimize the sizes of the HMM state observation densities with a genetic algorithm (GA). To the author's knowledge, the discriminative optimization for DGMS accomplishes discriminative training of discrete variables that is first proposed. We found the proposed framework is able to cover more multi-accent changes, thus reduce some performance loss in pruned beam search, without increasing the model size of the original acoustic model set. Evaluation on three typical Chinese accents, Chuan, Yue and Wu, shows that our approach outperforms traditional acoustic model reconstruction techniques with a syllable error rate reduction of 8.0%, 5.5% and 5.0%, respectively, while maintaining a good performance on standard Putonghua speech.**

***Index Terms*—Accented speech recognition, accent-specific unit, dynamic Gaussian mixture selection (DGMS), genetic algorithm.**

I. Introduction

ACCENT variability is a significant factor for performance degradation for most state-of-the-art automatic speech recognition (ASR). This is particularly serious for Mandarin Chinese ASR systems with a large speaking population and a wide range of accent variations when they use the same written Chinese characters that are ideographic and independent from their pronunciations. Accented speech is often caused by some pronunciation differences between the speaker's first language or dialect, and that of the target standard speech. Such discrepancies can be either acoustical or phonological.

There are seven major dialects in China: Guanhua, Yue, Wu, Xiang, Gan, Kejia and Min, which can be further divided into more than 30 sub-categories [1]. As a large proportion of Chinese speakers learn Putonghua as a second language their pronunciations are inevitably influenced by their native dialects. Statistics show that over 79.6% of Putonghua speakers carry regional accents, and 44.0% of them speak strong accents [2]. Furthermore, multitude of accents is very common in Chinese speech [3]. As a result, ASR systems implemented for standard Putonghua cannot perform well on accented speech, especially when multiple accents are involved.

There are quite a few previous studies on accent analysis [4]–[6], classification [7]–[9], and detection [10]. In our current study, we focus our attention on accented speech recognition.

Conventional methods that handle accent variations focus on modeling acoustic and phonetic variations at different levels. For phonetic variations, phone set extension and augmented pronunciation dictionary are commonly adopted methods [11]. However, they usually increase lexical confusions [12], but do not lead to a significant performance improvement. The other approaches to reduce phonetic confusions include pronunciation modeling at either phone-level or HMM-level [13], [14]. For acoustic variations, the most straightforward way is to build acoustic models for each accent using a large amount of accented data [15]. On the other hand, for the case of multiple accents, an extra accent classification module is needed [16] to identify the exact accent the speaker has, and uses the acoustic models relevant to that accent to perform speech recognition. Another method is to apply maximum a posteriori (MAP) or maximum likelihood linear regression (MLLR) for acoustic model adaptation so as to fit the speech characteristics of certain accents [17]–[20]. A major weakness of the conventional adaptation techniques is that the parameters of the acoustic models undergo an irreversible change, making the models less capable of covering other accents and the standard language.

To address issues of above methods, an approach to state-level pronunciation modeling with model reconstruction was proposed to handle both the acoustic and phonetic changes for accented speech recognition [1], [3], [21], [22]. These methods are able to cover multi-accent changes as well as retrain the

Manuscript received February 19, 2013; revised May 15, 2013; accepted May 21, 2013. Date of publication May 29, 2013; date of current version July 22, 2013. This work was supported by the Natural Science Foundation of China (61272233), and a research grant from Nokia-Tsinghua Joint Funding in 2008-2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shinji Watanabe.

C. Zhang is with the Machine Intelligence Laboratory, Engineering Department, University of Cambridge, Cambridge CB2 1PZ, U.K. (e-mail: cz277@cam.ac.uk).

Y. Liu is with the Shenzhen Key Laboratory of Intelligent Media and Speech, PKU-HKUST Shenzhen Hong Kong Institution, Shenzhen 518057, China (e-mail: yi.liu@imsl.org.cn).

Y. Xia is with the Center for Speech and Language Technologies, Tsinghua University, Beijing 100084, China (e-mail: yqxia@tsinghua.edu.cn).

X. Wang is with the Beijing Language and Cultural University, Beijing 100083, China (e-mail: xwang.blcu@gmai.com).

C.-H. Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Digital Object Identifier 10.1109/TASL.2013.2265087

performance for standard speech, and have finer model resolution than phone-level or HMM-level pronunciation modeling approaches [1], [3], [21]–[23]. These methods usually increase the model size resulting in an inefficient use of the Gaussian mixture components and sacrifice model resolution [23]. This inevitably brings a serious performance loss to beam pruning in the ASR decoding process.

To address the above challenges, we propose reliable accent-specific unit (ASU) generation with discriminative dynamic Gaussian mixture selection (DGMS) in multi-accent speech recognition. In other words, we extend previous studies of applying state-level pronunciation modeling approach on multi-accent speech recognition task by reliable ASUs generation and discriminative DGMS. Our contributions are summarized as follows:

1) Time-aligned phone recognition is proposed to generate reliable ASUs precisely and efficiently. Being able to eliminate the frame mismatches, the reliable ASU[1] candidates together with their corresponding training samples can thus be obtained. Meanwhile, time-aligned phone recognition is helpful to model multi-accent variations accurately.
2) In order to improve model resolution of the statically reconstructed states degraded by acoustic model shift due to accents, we propose a dynamic Gaussian mixture selection (DGMS) strategy to construct a dynamic observation density by choosing a number of Gaussian components being most representative with and nearest to each input speech frame. Such a selected observation density size for each HMM state is set individually, and optimized discriminatively based on a minimum classification error (MCE) criterion [24], [25]. We also propose to solve the implied integer programming optimization problem efficiently using a genetic algorithm (GA) [26]. To the author's knowledge, optimizing DGMS by MCE based GA implements discrete-variable discriminative training for the first time, which operates on observation density sizes rather than means, variances, and weights by regular continuous-variable discriminative training [24], [25]. As a result, the proposed discriminative DGMS algorithm covers the potential multi-accent variations in speech with an enhanced robustness via improved model resolution, without increasing the model size, and therefore reduces the performance degradation caused by pruning errors in beam search based ASR decoding.

Moreover, there are some related studies on mixture and frame selection in the literature [27]–[29]. Conventional Gaussian mixture selection approach aims at improving the efficiency of using Gaussian component while frame selection is often adopted to increase system accuracy by abandoning frames with high confusions. Different from these approaches, our discriminative DGMS enhances the degree of matching between the testing frames and the Gaussian components, which is more close to what we would like to accomplish in online adaptation [30]. In addition, Gaussian pruning and weight normalization based frame-wise model re-estimation approach has been proposed recently for noise robust voice activity detection [31]. This approach selects an uncertain number of dominant Gaussian components for a testing frame, which form a dynamic Gaussian Mixture Model (GMM) for computing the likelihood of that frame. Both frame-wise model re-estimation and our proposed DGMS approaches have the idea of building a dynamic GMM to evaluate the testing frame by selecting the most suitable Gaussian components. On the other hand, the dynamic GMM model sizes associated to the HMM states of DGMS are optimized discriminatively by MCE based GA. In other word, DGMS uses a more complex and automatically learnt Gaussian component selection strategy, which is applicable to HMM based continuous density ASR systems.

The rest of this paper is organized as follows. In Section II, we discuss the generation of reliable ASUs. In Section III, acoustic model reconstruction together with DGMS is described. In Section IV, the discriminative DGMS framework with MCE based, GA-style integer programming optimization is elaborated. In Section V, a series of accented speech recognition experiments is presented. Finally we conclude our findings in Section VI.

II. Reliable Accent-Specific Unit Generation

A. Chuan, Yue and Wu Accents

Guanhua (Mandarin) is the largest dialect and is commonly spoken in the vast areas of north and southwest China while the other major dialects are often used in one or two provinces in the southeast, the detailed geographical distribution of the seven major dialects together with their sub-dialects could be found in Table II in [32]. Particularly, Putonghua (the standard Chinese) bases on the sub-dialects of northern Guanhuan.

Chuan, Yue (Cantonese) and Wu accents are investigated in this study, which would be of extensive use. The Chuan dialect is a sub-dialect of Guanhua in southwest China with more than 100 million users. And the Yue and Wu dialects are both major Chinese dialects from the most development areas in China. All these accents have very different pronunciations compared with standard Putonghua. For example, linguists have shown that 60% of the pronunciations between Yue and Putonghua are not even close to each other [33]. Many of them often regard Yue as a distinctive spoken language from Putonghua, in terms of phonological, lexical and syntactic structures. Phonologically, the pronunciation distinction of the same Chinese character in Yue and Putonghua is considered as different as that of the same morphological word in French and English [1].

Chinese is a monosyllabic language with each of its syllable constituted by an initial followed by a final [34]. An initial refers to a consonant that starts a syllable except for the six zero initials: _a, _o, _e, _i, _u, and _v, which actually represent no physical phone but make the monophthong syllables have logical initials [35]. A final is a vowel followed by 0 to 2 consonants/vowels. Conventional Chinese initial/final notations with their corresponding International Phonetic Alphabet (IPA) symbols in parentheses are adopted in this paper. In typical Chinese ASR systems, initials and finals are commonly used as subword

[1]The term "reliable ASU" in this paper is an abbreviation of "the ASU that is generated by the reliable ASU generation" rather than a different kind of ASU.

units to construct acoustic models. These units are different between the three dialects and Putonghua. There are 22 initials and 36 finals in Putonghua in contrast to 20/38, 20/53 and 27/49 initials/finals in Chuan, Yue and Wu, respectively [33]. The inventories of the initials/finals in the four languages are distinct. For example, compared to other languages, Yue has an additional velar nasal initial /ng/. In the meantime initials in Wu are usually divided into voiced and unvoiced groups but not for those in the other languages.

Consequently, many speakers from each dialect region have difficulty in pronouncing some Putonghua initials/finals. For instance, when a Yue dialect speaker tries to pronounce a Putonghua initial , 'zh'(/tʂ/) the phone made may lie between 'zh'(/tʂ/) and 'z'(/ʦ/), causing an accent variation. An ASU is widely used to represent an accent variation whereas a diversity of multi-accent variations can be represented using different sets of ASUs [3].

B. Reliable Accent-Specific Unit Generation

In ASR, an accent variation is an erroneous pronunciation of a canonical phone, B, into an alternative, say S, due to the effect of an accent, causing a misrecognition error. The phone S is called an ASU of the phone B. Conventional methods for generating ASU candidates extract phone pairs from the alignment of the canonical and alternative transcriptions [3]. The canonical transcriptions are manually labeled, and the alternative transcriptions could be generated automatically by free grammar phone recognition that produces substitution (transfer), insertion and deletion (epenthesis) errors [36]. There is always a large portion of insertions and deletions in alternative transcriptions due to the poor performance of the pre-trained standard speech acoustic models on accented data which causes frame mismatches to substitutions. In our experiment, there are more than 2 insertions and 2 deletions for each Yue accent utterance on average while there is actually rare initial/final level epenthesis in Chinese according to consensus in linguistics [34]. Fig. 1 is a real example to show how frame mismatches impact on ASU generation. For ASU candidates generate by traditional methods, the canonical and alternative pairs of phones, '_i'→'j'(/ʨ/), 'iu'(/iəəu/)→ 'ie'(/iɛ̟/), '_u' →'m'(/m/), 'n'(/n/)→'_i', and 'ian'(/iɛn/)→ 've'(/yɛ̟/), belong to different audio slices and are not representative accent variations, that results in unreliable ASU.

We propose time-aligned phone recognition to eliminate insertion and deletion errors. ASR performs 1-best decoding based on the exact time boundaries of each phoneme obtained from a forced alignment [36] by imposing the two following selection principles: 1) the number of phone in the alternative result should be the same as that of the canonical transcription; 2) each phone in the selected alternative should be of the same duration as its corresponding phone in the canonical transcription. The implementation could be as simple as adding an operation of filtering all the hypotheses with these two selection rules before ranking them according to the original criterion (e.g., maximum a posterior criterion for the conventional Viterbi decoding). A more efficient implementation is to build a decoding network with the number of nodes in every possible path and the number of frames associated to every node (associates to an HMM definition [36]) in that network constrained by the number of segments and their durations in the forced aligned reference transcriptions, separately. Then the conventional network based searching and result selection could be applied to this network to fulfill this time-aligned phone recognition. In other words, time-aligned phone recognition performs context-dependent phone classification rather than recognition. This time-aligned phone recognition is different from slicing the data into individual initial/finals and do segment-by-segment recognition since context-dependent models could be applied in time-aligned phone recognition, which captures more accurate accent changes.

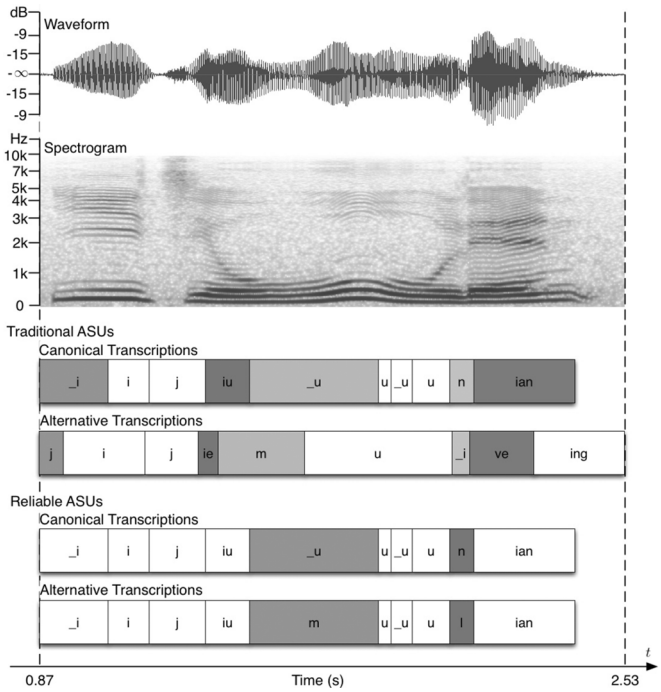


Fig. 1. A real example illustrates that frame mismatches cause inaccurate ASU instances in contrast to reliable ASU instances. The vertical axis is time. Each rectangle in a bar represents a phone with its length is the relevant phone duration. The results on the upper and lower parts in each group of bars are canonical and alternative transcriptions, respectively.

The procedure to generate reliable ASUs is illustrated in Fig. 2 and explained as follows:

1) *Acquiring canonical transcriptions with time:* a forced alignment is performed on phone-level canonical transcriptions with pre-trained acoustic models, and the canonical transcriptions containing all phone durations are obtained [36].
2) *Obtaining alternative transcriptions:* with phone durations obtained from the canonical transcriptions with time, the alternative transcriptions are generated by time-aligned phone recognition.
3) *Generating reliable ASU candidates:* reliable ASUs are those misclassified phones obtained by comparing the canonical and alternative transcriptions.
4) *Selecting reliable ASUs:* reliable ASU candidates include those with real accent variations as well as errors from data and recognizer confusions [1]. Hence, reliable ASUs are

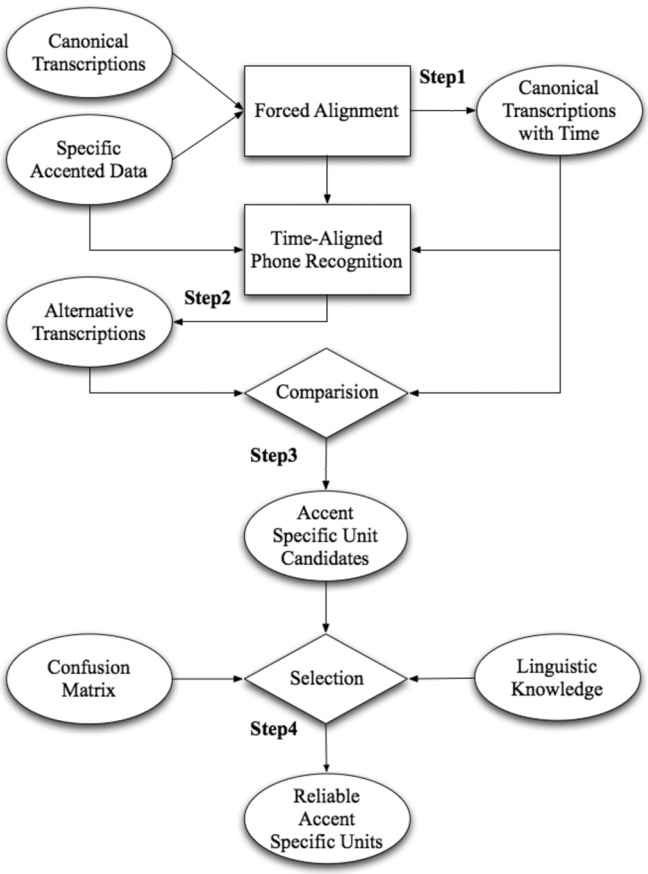


Fig. 2. Flow-chart for generating reliable accent-specific units.

manually selected from the candidates in reference with linguistic knowledge and a phone confusion matrix [1].

The selection strategy in Step 4 includes the following steps: (i) *removing language-inherent confusion*: for example, 'i2'(/ɭ/)→'i1'(/ɻ/) may not be a typical accent change in any of the three dialects if it takes a high confusion probability in Putonghua. It is then treated as a language-inherent confusion and removed; (ii) *replacing pronunciations of suspicious alternatives with their inherent confusions:* for instance, 'an'(/an/) → 'ai'(/aɪ/) may not be a typical Chuan variation. But Chuan speakers tend to pronounce 'an'(/an/) as /ae/, which is not a valid phone in Putonghua. Since the most similar final to /ae/ is 'ai'(/aɪ/), we then consider 'an'(/an/) → 'ai'(/aɪ/) as an accent variation; and (iii) *removing errors from either data or recognizers:* for example, the retroflex affricative 'zh'(/tʂ/) is not in any of the three dialects. 'z'(/ʦ/) →'zh'(/tʂ/) does not coincide with any linguistic knowledge either and has much fewer instances than 'zh'(/tʂ/)→'z'(/ʦ/). As a result, 'z'(/ʦ/)→'zh'(/tʂ/) is removed.

Remarkably, we can use the pre-trained acoustic phone models to obtain duration information associated with the conventional transcriptions. It is intuitive that acoustic models involving accent pronunciations generate more accurate segmentations on accented data. To show how much do accent changes affect the segment durations obtained by forced alignment, we mixed the development sets of Chuan, Yue, and Wu accents to adapt the pre-trained models using the MAP approach [3], [17]. Comparing the durations of the same phone generated by the pre-trained and adapted models, if their start or end time has a difference of more than 3 frames, the duration by the pre-trained models is supposed to be inaccurate. The threshold was set to 3 due to the well-known noisy nature of the phone boundaries. Our results obtained by the pre-trained models showed only 4.0%, 4.6%, and 3.0% of the phone segments resulted in inaccurate durations in the development sets of the Chuan, Yue, and Wu accents, respectively. These high duration accuracies can be attributed to the fact that the highly likely paths during the forced alignment procedure are mostly restricted to narrow regions. Therefore these results indicated that the pre-trained models could be used to take the place of the adapted models to acquire the durations of the accented phones. This fact simplifies the procedure for generating reliable ASUs.

Time-aligned phone recognition eliminates the frame mismatch problem and captures accurate accent variations. In Fig. 1, by reliable ASU generation, "_i', 'iu'(/iəu /), and 'ian'(/iɛn/) are correctly recognized phones rather than substitutions; '_u' and 'u'(/u/) are not deleted phones; the generated ASU candidates are then '_u' → 'm'(/m/) and 'n'(/n/) → 'l'(/l/) rather than the five candidates obtained by traditional methods. More specifically, compared to the traditional method that aligns canonical and alternative transcriptions by dynamic programming and edit distances [37], our method captures more accent changes (e.g.,'ch'(/tʂʰ/) → 's'(/s/) in Chuan accent). Furthermore, time-aligned phone recognition generates more reliable ASU candidates (e.g., 38 instances of 'un'(/uən/)→'en'(/ən/) in Chuan accent do no meet linguistic knowledge in our method whereas there are 61 such instances in the traditional method), which would be used as training samples for ASU models. Moreover, the selection is easier since there are always less recognition errors in time-aligned phone classification. With reliable ASUs and their corresponding training samples, the process results in accurate ASU model generation which plays an important role in subsequent acoustic model reconstruction.

It is also noteworthy that although there are similar ASUs in different accents (e.g., 'zh'(/tʂ/)→'z'(/ʦ/)), the tendency of the represented accent changes and their corresponding acoustic parameters are distinct [3]. In addition, speakers who have lived in more than one dialectal region tend to have mixed accents. For example, in the extreme case, pronunciation of 'zh'(/tʂ/) from such speakers can distribute over the entire range between 'zh'(/tʂ/) and 'z'(/ʦ/) [3]. Therefore, handling the multitude of accents requires more flexible acoustic models to be discussed in the next section.

III. Dynamic Gaussian Mixture Selection With Acoustic Model Reconstruction

We build triphone acoustic models for each set of reliable ASUs, and merge those Gaussian components borrowed from such models into the set of pre-trained acoustic models to cover accent changes, through acoustic model reconstruction [1]. A key objective is to statically reconstruct the state observation densities for a subset of the selected tied-states, and to increase the model robustness to handle multiple accent changes [3]. This is called an augmented HMM set.

Due to the problems caused by the increased model size and model resolution loss in the augmented HMM set, we propose

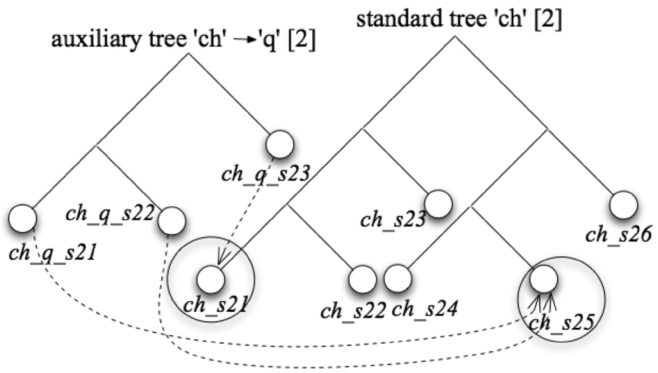


Fig. 3. Decision tree merging for ASU 'ch'→'q', where 'ch' is the canonical phone, and 'q' an alternative phone.

to use DGMS to improve the model reconstruction process by dynamically adjusting the statically reconstructed HMM state densities based on the acoustic similarities between each input speech frame and all the Gaussian components.

A. Acoustic Model Reconstruction for Multiple Accents

In the current ASR systems, words are presented by the concatenation of subword units (e.g., phones). The sequence of decoded units is obtained with maximum likelihood decoding as follows:

$$\hat{B} = \arg\max_{B} P(O|B)P(B), \tag{1}$$

where $O = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_N$ denotes the input frame sequence, N the frame number of that sequence, and $B = b_1, b_2, \ldots, b_N$ the canonical phone sequence corresponding to those frames. $P(O|B)$ represents the acoustic model probability, and $P(B)$ the language model probability that we will not consider in this study. Due to the effect of accents, some standard units can be pronounced incorrectly. Suppose $S = s_1, s_2, \ldots, s_N$ is one possible alternative pronunciation sequence, (1) needs to be rewritten to take accents into consideration as follows:

$$\hat{B} = \arg\max_{B} \left[P(B) \sum_{S} P(O|B,S)P(S|B) \right]. \tag{2}$$

In (2), $P(O|B,S)$ represents the acoustic model taking an alternative unit sequence S into account, and $P(S|B)$ the pronunciation model [38]. Both acoustic models $P(O|B)$ and $P(O|S)$ (if accented data and alternative transcriptions are available) are sub-optimal when both standard and accented speech would be met. We use the optimal model $P(O|B,S)$, which can be factorized in into successive contributions:

$$P(O|B,S) = \prod_{i=1}^{N} p(\mathbf{o}_i|b_i, s_i), \tag{3}$$

where $\mathbf{o}_i$ is the speech frame corresponding to a canonical phone b_i and its alternative phone s_i. Model $p(\mathbf{o}_i|b_i, s_i)$ is the likelihood of frame $\mathbf{o}_i$ given the relevant canonical phone b_i and alternative phone s_i [38].

In this paper, a decision tree-based tied-state triphone model is adopted [36]. We build a triphone model for each reliable ASU [1]. Decision trees for ASUs are named auxiliary trees in contrast to those for the pre-trained models that are called standard trees. Since each leaf node of a decision tree represents a tied-state, acoustic model reconstruction that borrows the accented Gaussian components from the ASU models to enlarge the original observation densities in the pre-trained HMM states is equivalent to merging the leaf nodes of the auxiliary tree into the standard tree leaf nodes. A leaf node of an auxiliary tree is merged into one leaf node on the standard tree relevant to its canonical phone, if these two nodes have the minimum Mahalanobis distance [1], as shown in Fig. 3. The above decision tree merge procedure is similar to the approaches used in cross language mapping [38]–[40].

The new output distribution $p'(\mathbf{o}|b)$ of the merged node for this statically reconstructed tied-state is defined as follows:

$$p'(\mathbf{o}|b) = \lambda p(\mathbf{o}|b) + (1-\lambda) \sum_{i=1}^{V} p(\mathbf{o}|b, v_i)p(v_i|b), \tag{4}$$

where $p(\mathbf{o}|b)$ denotes the output density of the pre-trained model, and λ is determined as the probability that the canonical phone is correctly recognized [1], and V is the total number of the merged nodes from the auxiliary trees. b and v_i are the canonical phone and alternative phone of an ASU. $p(v_i|b)$ is the confusion probability between b and v_i, and can be estimated from a confusion matrix [3]. After acoustic model reconstruction, Eq. (2) becomes

$$\hat{B} = \arg\max_{B} P'(O|B)P(B), \tag{5}$$

where $P'(O|B)$ is defined in (4).

B. Dynamic Gaussian Mixture Selection

The solid line in Fig. 4(a) illustrates the output observation density of a tied-state in pre-trained acoustic models. For an augmented HMM state, its statically reconstructed observation density is enlarged with the borrowed Gaussian components that extend its coverage to handle accent variations, as shown by the solid line in Fig. 4(b). Nevertheless, acoustic model reconstruction can significantly increase model size. For instance, in our experiment, 6,620 accented Gaussian components were merged into 546 standard tied-states. The state with the largest density size borrowed 120 accented Gaussian components belonging to various accent changes and placed them at different parts of its distribution. Therefore, the likelihoods of the frames located nearby the center of the augmented density decrease comparing to those by the pre-trained models, as illustrated by the center parts of the solid lines in Fig. 4(a) and (b). This likelihood reduction would result in confusions between the augmented models and its competing models, which causes model resolution loss.

The problem caused by model resolution degradation was alleviated by DGMS without making any changes to the acoustic models. This is achieved by selecting suitable Gaussian components to construct a dynamic observation density for each input speech frame according to a k-nearest principle. That is, k Gaussian components being nearest to the current frame are selected to customize a new output state observation density to calculate its acoustic log-likelihood. Considering different variances of the various Gaussian components, Mahalanobis

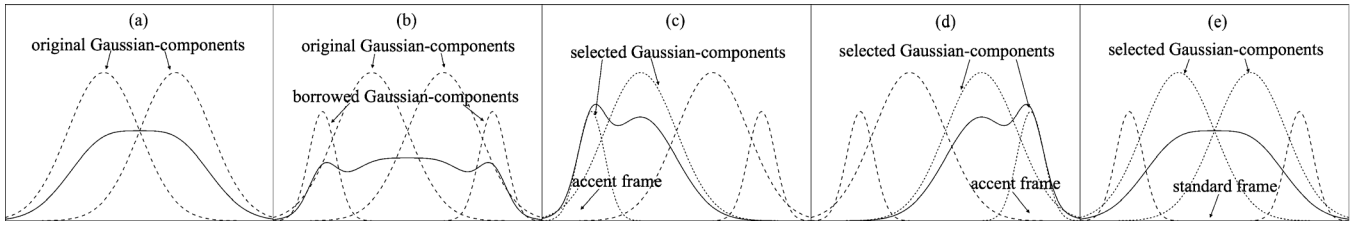


Fig. 4. Output densities of acoustic model reconstruction and DGMS. The solid and dashed lines draw the output densities and their constituent Gaussian components, respectively. (a) is the output density of pre-trained acoustic models by standard speech; (b) is the density of a augmented acoustic model obtained by static reconstruction with Gaussian components borrowed from auxiliary decision trees; (c) and (d) show the dynamic densities for different accented frames; (e) illustrates a dynamic density for standard frame.

distance is used to measure the distance between a frame and a Gaussian component since it coincides with the asymmetric property of acoustic confusions in accented speech [1], [4]. Hence, this principle can be presented as follows.

Note $b_r(\mathbf{o}) = \sum_{m=1}^{M} w_m \mathrm{N}(\mathbf{o}|\mu_m, \Sigma_m)$ is the density for a statically reconstructed state r (also noted as $p'(\mathbf{o}|b)$ in Section III-A). Suppose $N'_1, N'_2, \ldots, N'_k$ are the k components nearest to $\mathbf{o}$ among all the M Gaussian components, $N'_m = \mathrm{N}(\mathbf{o}|\mu'_m, \Sigma'_m)$, then a dynamic observation density for speech frame $\mathbf{o}$ on state r is:

$$b'_r(\mathbf{o}) = \sum_{m=1}^{k} w''_m \mathrm{N}(\mathbf{o}|\mu'_m, \Sigma'_m), \tag{6}$$

where $w''_m = w'_m / \sum_{m=1}^{k} w'_m$ is the normalized weight of the selected Gaussian components in dynamical observation density.

For an accent frame located at the boundary of the density, the principle of the k-nearest neighbor selects k Gaussian components being nearest to the current frame, that are the most representative Gaussian components for the relevant accent change. Consequently, the obtained dynamic densities have better model representation ability for accent changes as illustrated by the solid lines in the left part and right part of Fig. 4(c) and (d), separately, which reduces the performance degradation in state pruning during beam search in decoding. Meanwhile as shown in Fig. 4(e), for an acoustic sample of standard speech located at the center of the statically reconstructed density, its dynamic output density drawn by the solid line in the center part of the figure would still be similar to that before model reconstruction, and therefore retain its covering ability for standard speech.

It is remarkable that the nature of DGMS approach is to select a Gaussian mixture for an input frame rather than select individual Gaussian components, due to the adoption of the k-nearest principle. For a specified k, there are C_M^k possible combinations of the selected Gaussian components in the reconstructed HMM state. A combination of the selected Gaussian components forms a Gaussian mixture, which could be used as a dynamic observation density. Therefore, the approach is named as dynamic Gaussian mixture selection as an indication of this nature.

In addition, k is pre-specified distinctively for different HMM states that provides sound flexibility to fit the diversified accent variation. This gives many flexible parameters to optimize in statically reconstructed HMM states for DGMS, which are referred to as the parameter vector. Challenges in finding the optimal parameter vector include: (i) how to evaluate a parameter vector; and (ii) how to find the optimum.

IV. MCE Based Discriminative Discrete Variable Optimization for Dynamic Gaussian Mixture Selection

To estimate the parameter vector (namely an integer-valued vector with every of its element associated to the number of selected Gaussian components in a reconstructed state) for DGMS precisely, we choose the MCE criterion that directly optimizes on segment error rate [24], [25]. The selected Gaussian mixture model sizes are discrete variables for the MCE loss function. Therefore, the optimal parameter vector cannot be solved by conventional optimization techniques used for discriminative training of continuous variables (e.g., mean, variance, etc.) [24], [25], since the derivatives for the Gaussian mixture model sizes do not exist. Parameter optimization for DGMS is actually an integer programming problem, so we use GA to solve it efficiently [26].

A. MCE Criterion for Discriminative DGMS

MCE is widely used to minimize errors in the training set. Its effectiveness has been proved in discriminative training of HMMs for ASR [24], [25], [41], [42]. In this study, we utilize MCE to estimate the parameter vector in 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 a discriminant function $g_j(O, \Lambda, c)$ [24], [25], which evaluates the acoustic log-likelihood for an output string S_j scored by DGMS with its parameter vector, c. A misclassification measure $d_i(O, \Lambda, c)$ is used to evaluate the acoustic difference between the canonical string S_i and its alternative strings.

$$d_i(O, \Lambda, c) = -g_i(O, \Lambda, c) + \max_{j, j \neq i} g_j(O, \Lambda, c). \tag{7}$$

For the convenience, we call the alternative string that has the largest score as the best incorrect string. Therefore, the loss function for MCE can be defined as follows.

$$l(d_i(O, \Lambda, c)) = \frac{1}{1 + e^{-\alpha \cdot d_i(O, \Lambda, c)}}. \tag{8}$$

We use the MCE loss function as the objective function in optimizing the parameters of DGMS. When the optimal value in (8) is obtained, DGMS minimizes the empirical risk by reducing the training set errors, including those caused by accents and increasing the covering ability for such variations.

TABLE I
DATA SETS SEPARATION IN EXPERIMENTS

ID	DevC	TestC	DevY	TestY	DevW	TestW	TestP
Duration	6.5h	4.3h	6.1h	3.5h	6.6h	3.8h	3.9h
Syllable Number	51,907	33,847	51,341	31,191	52,584	29,888	23,158
Speaker Number	20	20	20	20	20	20	10
Utterance Number	3,205	2,000	3,091	2,000	3,471	2,000	2,000
Type	Chuan accent		Yue accent		Wu accent		standard Putonghua

The canonical and alternative transcriptions are generated by forced alignment with free grammar phone recognition, keeping track of the full state alignment [36]. To obtain the best incorrect strings for a given vector c, ideally we need to decode or rescore the pre-generated lattices with Λ and c. However, since a large number of parameter vectors would be examined, even if we cache the likelihoods for every frame on every Gaussian component to speed up the likelihoods calculation, the computation cost for either decoding or lattice rescoring is not affordable. We acquire the best incorrect string as the phone sequence that score the largest among n fixed strings, which were generated as the n-best outputs with the reconstructed acoustic models Λ and rescored with current parameter vector c, for the purpose of reducing the computation cost. In order word, we use the pre-generated n-best outputs to approximate the lattice for vector c. Therefore, according to (7) and (8), when the recognition result for a sequence is correct, $d_j(O, \Lambda, c) < 0$. Otherwise $d_j(O, \Lambda, c) > 0$. Hence, the optimization is to find the optimal vector that results in the smallest value in (8).

It is worth noting that since our formulation is a discrete variable optimization problem, there are no continuous variable derivatives in the objective function. Therefore, the reason why we use the MCE loss function as the optimization objective rather than directly count the erroneous frames does not lie in its smooth functional form, but in the followings: 1) the MCE loss function allows different erroneous frames to have different impacts to the objective function, and trade off these impacts within an utterance to make the optimization on segment level instead of individual frames; 2) we are able to control the optimization using varying values of α to assign different weights to training utterances with different correct/incorrect degrees [41]. When α is small (e.g., 0.01), training tokens of many utterances are mapped to fall into the linear region of (8), that increases the separation between correct and incorrect hypotheses as well as the generalization ability on unseen data; when α grows large enough (e.g., 2.0), the MCE loss function counts the incorrectly recognized utterances. A more detailed discussion about the effectiveness of different α can be found in [41].

B. Genetic Algorithm

Let R denote the number of statically reconstructed states, the rth such state has n_r Gaussian components. Therefore, the number of possible parameter vector candidates for DGMS is $\prod_{r=1}^{R} n_r$, which grows exponentially as R increases. As there are usually hundreds of statically reconstructed states, and every such state has tens of borrowed Gaussian components, traversing every possible parameter vector to obtain the optimum is computationally infeasible.

GA, which is a "search for solutions" algorithm mimics the "survival of the fittest" process of natural evolution, is able to find the optimal solution by examining over only a small fraction of possible candidates [26]. Viewing a parameter vector as an individual chromosome in evolution, the optimization problem for finding the optimal parameters can be regarded as selecting the fittest one from every chromosome appeared in evolution, and the optimization can be efficiently handled by GA. Therefore, a chromosome c is constituted by R positive integers corresponding to the Gaussian mixture model size k for the R statically reconstructed states in DGMS. Every integer ranges from a pre-specified minimum to the maximum model size in its relevant statically reconstructed state.

We use the MCE loss function 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 solves the optimal parameter vector for DGMS as follows:

1) Randomly generate N chromosomes as the initial population;
2) Compute a fitness function $f(c)$ for each chromosome c;
3) Select C chromosomes randomly from the population, C is even. The selection probability is calculated as,
$$P(c) = \frac{1}{f(c)} \Bigg/ \left(\sum_{c'} \frac{1}{f(c')} \right), \tag{9}$$
where c' refers to every chromosome in current population. The "roulette-wheel sampling" method is used to make the selection [26]. It generates a random positive number less equal than 1, and successively sums $P(c)$ for every c in the current population until a chromosome c^* that causes the sum to exceed the random sample size, then c^* is selected.
4) Reproduce the C chromosomes by sequentially dividing them into $C/2$ pairs, then "one-point crossover" is used for the reproduction of each pair [26]. That is, selecting a random crossover point for each pair, and swapping integers beyond that point in either chromosome between the parents. This step generates C children.
5) Merge the children chromosomes generated in step 4 with the unselected ones in step 3, and form a new population. Generate a random probability for each chromosome at each locus in the new population. 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 with the new population.
7) Repeat steps 2–6, if no c satisfies $f(c) \approx 0$ and the maximum generation number is not reached.

TABLE II
DETAILS OF THE ASUS AND THEIR CORRESPONDING AUXILIARY DECISION TREES

System ID	System 2			System 3		
Dataset	DevC	DevY	DevW	DevC	DevY	DevW
ASU Number	160	187	166	165	191	166
Auxiliary Decision Tree Number	480	561	498	495	573	498
Auxiliary Tied-State Number	531	582	569	517	605	533

V. RECOGNITION EXPERIMENTS

A. Dataset

The 863 regional accent speech corpus [43] was used in our experiments to evaluate our method on three typical accents—Chuan, Yue, and Wu. This database is the largest and most commonly used one in Chinese accented speech recognition [1]. The training set is consisted of 25,920 utterances from 100 speakers with 51.5 hour of standard Putonghua speech. This set was used to build the baseline Putonghua system. All speech data were sampled at 16 kHz and with a 16-bit precision. More details are listed in Table I. To prove the effectiveness of the proposed approach, we selected the development sets (DevC, DevY and DevW) and testing sets (TestC, TestY, and TestW) as shown in Table I for each accent with the speakers representing strong accents based on their recording records.

B. Recognition Systems

Five systems were tested, which are described as follows.

System 1: The Baseline system. The acoustic model was trained using the training set. It is built on HTK decision tree based state tying procedures with 3,000 tied-states triphone models and 12 Gaussian components per state [36]. The HMM topology is 3-state, left-to-right without skips. The acoustic features are $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. 28 initials and 36 finals, including 6 zero-initials, in standard Chinese were used as the subword units for building HMMs.

System 2: Augmented HMMs with traditional ASU. Traditional ASUs were extracted from the alignment on the utterances in DevC, DevY and DevW, individually, which were generated by Flexible Alignment Tool [37]. The auxiliary decision trees are also built by HTK decision tree based state tying procedures with 4 Gaussian components per state. Details of the auxiliary decision trees for the tradition ASUs and their relevant tied-states are listed in Table II. The auxiliary tied-states were merged into the same acoustic model as System 1. The statically reconstructed model included 42,728 Gaussian components to be used in System 2. This system was built in order to show the effectiveness of conventional ASUs,

System 3: Augmented HMMs with reliable ASU. We list the details of the reliable ASUs, their corresponding decision trees, and tied-states for the three accents respectively in Table II. The reliable ASUs were generated for each accent individually according to the procedure described in Fig. 2. We constructed auxiliary trees for the reliable ASUs and merged tied-states into the baseline acoustic models through acoustic model reconstruction. The augmented acoustic models in System 3 include 42,620 Gaussian components with 14.2 mixtures per state on average.

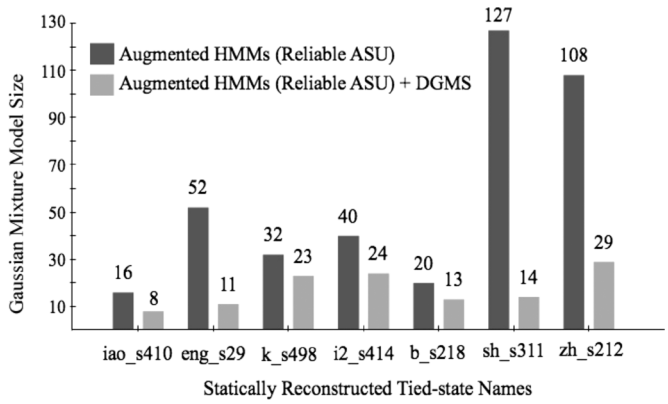


Fig. 5. Gaussian mixture model sizes in state observation densities for some representative tied-states with/without discriminative DGMS.

System 4: System 3 with discriminative DGMS. Our System 3 included 546 statically reconstructed tied-states. Therefore, DGMS requires 546 parameters for model sizes. To optimize these discrete variables, α for the MCE criterion was set to 0.01; and the population size N, reproduction size C, and mutation probability for GA were set to 400, 200, and 0.3, respectively. The 5-best hypotheses generated by System 3 were used to approximate the exact alternative transcriptions in GA. We optimized the DMGS parameter vector with utterances from the DevC, DevY, and DevW sets. The optimization was forced to stop at a maximum of 1,200 iterations, and its objective function value in (8) was converging and reduced from 0.86 to 0.62. The obtained solution was used as the parameters for discriminative DGMS in all experiments. Some selected sizes in our optimized vector for some representative tied-states are compared in Fig. 5. Clearly these model sizes are smaller with DGMS than those used in System 3 without DGMS.

System 5: System 1 acoustic models being MAP-adapted to fit the acoustic characteristics of Chuan, Yue, and Wu accents using a mixture of data from the DevC, DevY, and DevW sets. This system was built as a control group for comparing our approaches with conventional accent adaptation approaches.

C. Results and Discussions

All results in syllable error rate (SER) with the five systems discussed earlier are listed in Table III with a free grammar for recognizing 410 Chinese syllables, in which decoding any syllable can follow any syllable with equal probability. The reason these studies evaluated the performance without language models is to detach the influence of higher-level information from the issue objective of obtaining the real acoustic model improvement. Compared to the baseline (System 1), the augmented HMM system with traditional ASUs (System 2) yielded a SER reduction on every accented testing set. The reason lies in the fact that the borrowed accented Gaussian

TABLE III
LOWER SER FOR USING OUR APPROACH COMPARED TO USING TRADITIONAL ASU AND MAP ADAPTATION

ID	System	Syllable Error Rate (SER) %			
		TestP	TestC	TestY	TestW
1	**Baseline**	22.5	56.6	58.3	55.5
2	**Augmented HMMs (Traditional ASU)**	22.2 (-0.3)	49.3 (-7.3)	49.0 (-9.3)	46.9 (-8.6)
3	**Augmented HMMs (Reliable ASU)**	22.1 (-0.4)	47.1 (-9.5)	48.3 (-10.0)	46.4 (-9.1)
4	**Augmented HMMs (Reliable ASU) + DGMS**	22.2 (-0.3)	45.4 (-11.2)	46.4 (-11.9)	44.6 (-10.9)
5	**MAP adaptation with DevC, DevY and DevW**	30.6 (+8.1)	49.2 (-7.4)	48.3 (-10.0)	46.9 (-8.6)

components in the statically reconstructed tied-states adjusted the original distribution and enabled more Gaussian components at the boundaries to cover the confusing pronunciation of accent changes [1], [3]. Compare to System 2, System 3 gives a relative SER reduction of 4.5%, 1.5% and 1.1% on TestC, TestY and TestW, respectively. These results indicate that the reliable ASUs have better covering ability for accent variations than the traditional ASUs.

With discriminative DGMS, System 4 obtains a lower relative SER by 3.6%, 4.0% and 3.9% than System 3 on TestC, TestY and TestW, respectively. Discriminatively trained DGMS minimizes the classification errors in the training set and increases the coverage abilities for multi-accent variations. Meanwhile, our proposed DGMS method also maintained the recognition performance on Putonghua as we restricted the dynamic output densities with no less than 6 Gaussian components by DGMS, leading to retaining the pre-trained mixture distributions for standard speech. With the joint usage of our proposed reliable ASU and discriminative DGMS (System 4) achieves 8.0%, 5.5% and 5.0% lower relative SER than the traditional acoustic model reconstruction approach (System 2) for TestC, TestY and TestW, respectively. System 4 clearly achieved the best SER among all competing systems for multi-accent speech.

It is also noted from the two bottom rows in Table III that System 4 outperforms System 5 by 7.8%, 4.0% and 4.9% lower relative SER reduction. These results show: (i) explicitly and accurately modeling each accent change is better than adapting a model to fit accent changes for all accents together; and (ii) the discriminative MCE criterion tackles the classification errors directly which outperforms the generative MAP criterion. Meanwhile, System 4 does not degrade the system performance on standard Putonghua speech achieving an SER of 22.2%, whereas System 5 severely dropped the SER to 30.6%. For our setup MAP adaptation adjusts the acoustic model parameters to fit multi-accents, and makes them no longer well-fit for standard speech.

An example for using discriminative DGMS to reduce local model mismatch for the Yue accent is illustrated in Fig. 6. In this example, the initial 'zh'(/tʂ/) in syllable 'zhi'(/tʂl/) was misrecognized as 'z' when using the baseline and System 3. This is caused by the most typical Yue accent change between 'zh'(/tʂ/) and 'z'(/ʦ/). The 3 states of the misrecognized initial were presented from frames 215 to 230. The acoustic log-likelihoods for both the baseline and the augmented HMM system with reliable ASUs severely drop at around frame 222. In the augmented HMM system with reliable ASUs, the borrowed accented Gaussian components for 'zh'(/tʂ/) → 'z'(/ʦ/) is helpful to increase the acoustic log-likelihood at frame 222,

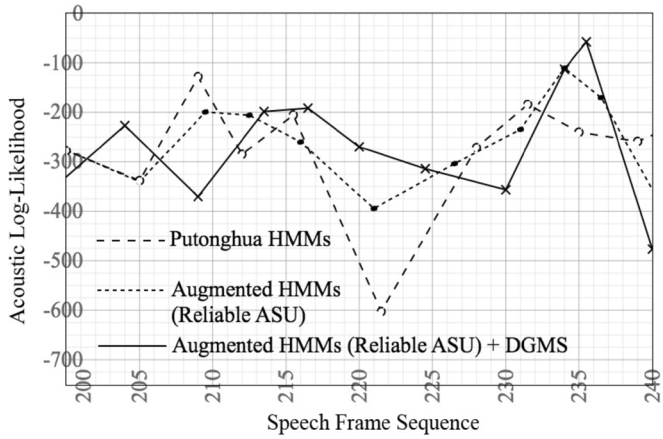


Fig. 6. An example of using discriminative DGMS to correct local model mismatch in Yue accent.

but not enough to restore the local mode mismatch. Moreover, with discriminative DGMS, the robustness for covering accent changes is improved as shown in the solid curve with the dynamic observation densities further increased the covering ability for 'zh'(/tʂ/) → 'z'(/ʦ/). As a result, System 4 reduces the degree of this local model mismatch and gives a correct recognition result.

Finally, we examine the effectiveness in using DGMS to avoid model resolution loss. We evaluate the resolution by

$$D_i(\Lambda) = -\sum_{O} d_i(O, \Lambda, c)/g_i(O, \Lambda, c), \tag{10}$$

where g_i is the acoustic log-likelihood with canonical state i; d_i is the same as in (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 Λ_0 denote the set of acoustic parameters of the augmented HMMs with traditional ASUs, $D_i(\Lambda_0)$ its resolution, Λ' a set of acoustic parameters, the relative model resolution improvement by Λ' can be defined as follows in (11) to measure the ratio of a discrimination power in terms of misclassification measures,

$$\mathrm{sgn}(D_i(\Lambda')) \cdot |D_i(\Lambda')/D_i(\Lambda_0)|. \tag{11}$$

The relative model resolution improvements as measure in (11) for some selective tied-states are illustrated in Fig. 7, which shows DGMS improved the relative model resolution, i.e., enhancing the separation between the canonical states and their competing alternative states. The larger the model resolution

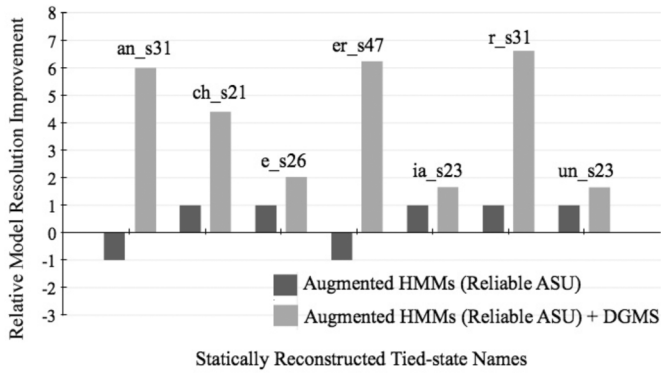


Fig. 7. Relative model resolution for representative tied-states with/without discriminative DGMS evaluated on Yue accent.

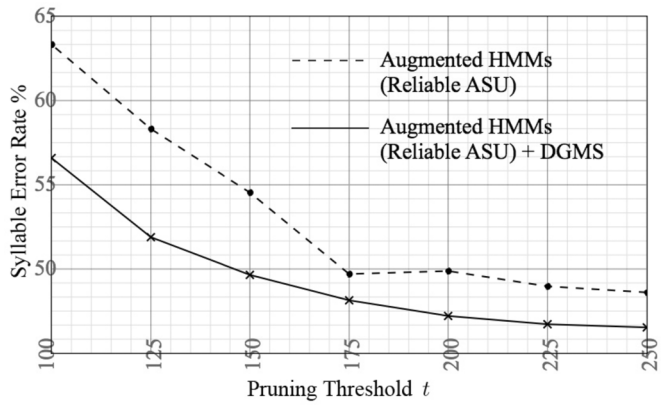


Fig. 8. SERs for augmented HMMs (reliable ASUs) with and without DGMS, when the pruning threshold varies from 100.0 to 250.0 in testing Yue speech.

improvement, the better the model separation, which also indicates enhanced separation between the target model and its competing models. Clearly, the model resolution is better with DGMS than that without DGMS.

This effect can also be reflected in beam search where a pruning threshold is set so that the models whose maximum acoustic scores fall below it are deactivated in decoding. By having a good model resolution the search paths reserved by DGMS are more accurate than those saved without DGMS, relieving the performance degradation in pruned search caused by acoustic model reconstruction, as we demonstrate in the two sets of curves in Fig. 8. With DGMS, the SER is always lower than that without DGMS. Furthermore the system performance degradation with smaller thresholds is more severe without DGMS than those with DGMS. Obviously, the SER difference between the two curves in Fig. 8 widens when the threshold value decreases indicating an enhanced robustness with a better model resolution for DGMS than that without DGMS. Remarkably, the power of DGMS comes with a price of increasing the computation complexity of decoding. Since Mahalanobis distance can be regarded as a part of the likelihood, the computing of the distances will yield no additional computation cost. The extra computations arise from sorting the Gaussian components based on their distances to current input frame. Therefore, for a set of augmented HMMs with R reconstructed states, R sorting would be performed for a testing frame. If quicksort is adopted, the computation complexity for sorting M Gaussian components is $O(M \log M)$.

VI. Conclusion

Unexpected acoustic variations occur constantly at run time in automatic speech recognition. In this paper, we propose to use time-aligned phone recognition to efficiently generate reliable accent-specific units. It is shown that reliable ASUs are able to capture accent changes accurately and explicitly at both the phonetic and acoustic levels. We also propose DGMS (dynamic Gaussian mixture selection) to handle multi-accent speech variation. When dynamically restructuring a pre-trained set of Gaussian mixture models at each input speech frame to cover the unexpected variations at run time, we show that discriminative DGMS, trained with genetic algorithm based on the minimum classification error criterion, improves model robustness and enhances model resolution for accented speech recognition, with an increasing acoustic covering ability for accent changes as well as reducing performance degradation in pruned beam search, and achieves discrete-variable discriminative training through integer programming for the first time known by the authors. Experimental results show that the combination of our proposed approaches yield a relative syllable error rate reduction of 8.0%, 5.5% and 5.0% on Chuan, Yue and Wu accents, respectively, when compared with the traditional acoustic model reconstruction techniques. Over conventional MAP adaptation, our approach achieves an SER reduction of 7.8%, 4.0% and 4.9% on the three accents, individually, with no accuracy impact on standard Putonghua.

Acknowledgment

The authors would like to thank Prof. Thomas Fang Zheng, Dr. Jesper Olsen, and Dr. Jilei Tian at Tsinghua University and Nokia Research Center respectively, for many useful discussions. The authors also appreciate the anonymous reviewers, for their valuable comments and suggestions helping us to improve the quality of this paper.

References

[1] P. Fung and Y. Liu, "Effects and modeling of phonetic and acoustic confusions in accented speech," *J. Acoust. Soc. Amer.*, vol. 118, no. 4, pp. 3279–3293, Nov. 2005.

[2] "Leading group office of survey of language use in China," in *Survey of Language Use in China*. Beijing, China: Yu Wen Press, 2006, (in Chinese).

[3] Y. Liu and P. Fung, "Multi-accent Chinese speech recognition," in *Proc. Interspeech*, 2006, pp. 1887–1990.

[4] M. Y. Tsai and L. S. Lee, "Pronunciation variation analysis based on acoustic and phonemic distance measures with application examples on Mandarin Chinese," in *Proc. IEEE Workshop Autom. Speech Recogn. Understand.*, 2003, pp. 117–122.

[5] S. Y. Chang, "A syllable, articulatory-feature, and stress-accent model of speech recognition," Ph.D. dissertation, Int. Comput. Sci. Inst. , Berkeley, CA, USA, 2002.

[6] S. Vaseghi *et al.*, "Analysis of acoustic correlates of British, Australian and American accents," in *Proc. IEEE Workshop Autom. Speech Recogn. Understand.*, 2003, pp. 345–350.

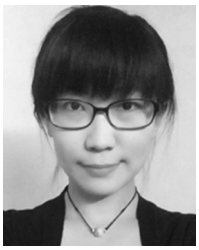
[7] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 836–839.

[8] K. Kumpf, "Automatic accent classification of foreign accented Australian English speech," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 1740–1743.
[9] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 634–646, Mar. 2006.
[10] Y.-L. Zheng *et al.*, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *Proc. Interspeech*, 2005, pp. 217–220.
[11] G.-H. Ding, "Phonetic confusion analysis and robust phone set generation for Shanghai-accented Mandarin speech recognition," in *Proc. Interspeech*, 2008, pp. 1129–1132.
[12] E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo, "A framework for predicting speech recognition errors," *Speech Commun.*, vol. 46, no. 2, pp. 153–170, Jun. 2005.
[13] E. Fosler-Lussier, "Dynamic pronunciation models for automatic speech recognition," Ph.D. dissertation, Int. Comput. Sci. Inst., Berkeley, CA, USA, 1999.
[14] T. Hain and P. C. Woodland, "Dynamic HMM selection for continuous speech recognition," in *Proc. Eurospeech*, 1999, pp. 1327–1330.
[15] V. Fisher *et al.*, "Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998.
[16] F. S. Richardson *et al.*, "Discriminative n-gram selection for dialect recognition," in *Proc. Interspeech*, 2009, pp. 192–195.
[17] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[18] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, Apr. 1995.
[19] Y.-R. Oh and H.-K. Kim, "MLLR/MAP adaptation using pronunciation variation for non-native speech recognition," in *Proc. IEEE Workshop Autom. Speech Recogn. Understand.*, Dec. 2009, pp. 216–221.
[20] P. Smit and M. Kurimo, "Using stacked transformations for recognizing foreign accented speech," in *Proc. Int. Conf. Audio, Speech, Signal Process.*, 2011, pp. 5008–5011.
[21] M. Saraclar *et al.*, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Comput. Speech Lang.*, vol. 14, pp. 137–160, Feb. 2000.
[22] M. Saraclar and S. Khudanpur, "Pronunciation change in conversational speech and its implications for automatic speech recognition," *Comput. Speech Lang.*, vol. 18, no. 4, pp. 375–395, Oct. 2004.
[23] Y. Liu and P. Fung, "State-dependent phonetic tied mixtures with pronunciation modeling for spontaneous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 351–364, Jul. 2004.
[24] B.-H. Juang *et al.*, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.
[25] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proc. IEEE*, vol. 86, no. 8, pp. 2345–2373, Aug. 2002.
[26] M. Mitchell, *An Introduction to Genetic Algorithm*. Cambridge, MA, USA: MIT Press, 1998.
[27] R.-Q. Huang and J. H. L. Hansen, "Unsupervised discriminative training with application to dialect classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2444–2453, Nov. 2007.
[28] Y. Lei and J. H. L. Hanzen, "Dialect classification via text-independent training and testing for Arabic, Spanish, Chinese," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 85–96, Jan. 2011.
[29] R. Chitturi and J. H. L. Hansen, "Dialect classification for online podcasts fusing acoustic and language based structural and semantic information," in *Proc. ACL-HLT*, Stroudsburg, PA, USA, 2008, pp. 21–24.
[30] C. Barras, S. Meignier, and J.-L. Gauvain, "Unsupervised online adaptation for speaker verification over the telephone," in *Proc. Odyssey*, 2004, pp. 157–160.
[31] M. Fujimoto, S. Watanabe, and T. Nakatani, "Frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection," *Speech Commun.*, vol. 54, no. 2, pp. 229–244, Feb. 2012.
[32] Y. Liu *et al.*, "HKUST/MTS: A very large scale Mandarin telephone speech corpus," in *Proc. Int. Symp. Chinese Spoken Lang. Process.*, 2006, pp. 724–735.
[33] J.-H. Yuan *et al., Survey of the Chinese Dialects*, 2nd ed. Beijing, China: Yu Wen Press, 2001, (in Chinese).
[34] T. Lin and L.-J. Wang, *A Course in Phonetics*. Beijing, China: Peking Univ. Press, 1992.
[35] J.-Y. Zhang *et al.*, "Improved context-dependent acoustic modeling for continuous Chinese speech recognition," in *Proc. Eurospeech*, 2001, pp. 1617–1620.
[36] S. Young *et al., The HTK Book*, 3.4 ed. Cambridge, U.K.: Entropic Cambridge Research Laboratory, 2009.
[37] P. Fung *et al.*, "Pronunciation modeling of Mandarin casual speech," in *The Johns Hopkins University Summer Workshop*, Baltimore, MD, USA, 2000, Final Report.
[38] Y. Liu and P. Fung, "Pronunciation modeling for spontaneous Mandarin speech recognition," *Int. J. Speech Technol.*, vol. 7, pp. 155–172, 2004.
[39] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proc. Interspeech*, 2009, pp. 528–531.
[40] J. Dines, L. Saheer, and H. Liang, "Speech recognition with speech synthesis models by marginalising over decision tree leaves," in *Proc. Interspeech*, 2009, pp. 1395–1398.
[41] E. McDermott and T. J. Hazen, "Minimum classification error training of landmark models for real-time continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 937–940.
[42] H. Watanabe *et al.*, "Minimum error classification with geometric margin control," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 2170–2173.
[43] A. Li *et al.*, "RASC863-A Chinese speech corpus with four regional accents," in *Proc. Oriental-COCOSDA*, 2004.

Chao Zhang (S'11) received the B.S. and M.S. (Hons) degree in computer science from Tsinghua University, Beijing, China, in 2009 and 2012. He is currently pursuing the Ph.D. degree at the University of Cambridge, supported by Cambridge International Scholarship. His research interests include automatic speech recognition and machine learning.

Yi Liu received the Ph.D. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology, Hong Kong in 2002. He is currently an Associate Professor of Shenzhen Key Laboratory of Intelligent Media and Speech, PKU-HKUST Shenzhen Hong Kong Institution. He was Research Associate in the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, from 2002 to 2006. He was an Associate Professor in Tsinghua University from 2007 to 2011. His research interests are spontaneous and conversational speech recognition, acoustic model and pronunciation model, accented speech recognition, speech and audio search. He is a member of IEEE and ISCA, respectively.

Yunqing Xia (M'04) is an Associate Professor in the Research Institute of Information Technology at Tsinghua University. He obtained his Ph.D. in computer science from the Institute of Computing Technology, Chinese Academy of Science in 2001. He worked in the Department of Computer Science at the University of Sheffield, UK from January 2003 to September 2004. From December 2004 to October 2006, he worked in the Department of System Engineering and Engineering Management at the Chinese University of Hong Kong. His research interests are natural language processing, text mining and information retrieval. He has been a standard member of IEEE Computer Science Society since 2004 and a professional member of ACM since 2007.



Xuan Wang received a B.A. in English language and literature from Beijing Language and Culture University, where she is currently completing an M.A. with a research focus on applied linguistics. Her current main research interest is in contrastive analysis on phonology and phonetic systems of English and Chinese (including mandarins and some regional dialects), which could be applied in Chinese ESL teaching and learning.

Chin-Hui Lee (S'78–M'82–SM'91–F'97) is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Dr. Lee received the B.S. degree in electrical engineering from National Taiwan University, Taipei, in 1973, the M.S. degree in engineering and applied science from Yale University, New Haven, in 1977, and the Ph.D. degree in electrical engineering with a minor in Statistics from University of Washington, Seattle, in 1981.

Dr. Lee started his professional career at Verbex Corporation, Bedford, MA, and was involved in research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, where he engaged in research and product development in speech coding, speech synthesis, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Between 1986 and 2001, he was with Bell Laboratories, Murray Hill, New Jersey, where he became a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. His research interests include multimedia communication, multimedia signal and information processing, speech and speaker recognition, speech and language modeling, spoken dialogue processing, adaptive and discriminative learning, biometric authentication, and information retrieval. From August 2001 to August 2002 he was a visiting professor at School of Computing, The National University of Singapore. In September 2002, he joined the Faculty Georgia Institute of Technology.

Prof. Lee has participated actively in professional societies. He is a member of the IEEE Signal Processing Society (SPS), Communication Society, and the International Speech Communication Association (ISCA). In 1991–1995, he was an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. During the same period, he served as a member of the ARPA Spoken Language Coordination Committee. In 1995–1998 he was a member of the Speech Processing Technical Committee and later became the chairman from 1997 to 1998. In 1996, he helped promote the SPS Multimedia Signal Processing Technical Committee in which he is a founding member.

Dr. Lee is a Fellow of the IEEE, and has published 400 papers and 30 patents. He received the SPS Senior Award in 1994 and the SPS Best Paper Award in 1997 and 1999, respectively. In 1997, he was awarded the prestigious Bell Labs President's Gold Award for his contributions to the Lucent Speech Processing Solutions product. Dr. Lee often gives seminal lectures to a wide international audience. In 2000, he was named one of the six Distinguished Lecturers by the IEEE Signal Processing Society. He was also named one of the two ISCA's inaugural Distinguished Lecturers in 2007–2008. He won the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition". In 2012 he was selected as an ISCA Fellow in 2012, and awarded the 2012 ISCA Medal in scientific.