RELIABLE ACCENT SPECIFIC UNIT GENERATION WITH DYNAMIC GAUSSIAN MIXTURE SELECTION FOR MULTI-ACCENT SPEECH RECOGNITION

Chao Zhang[⋆†]*, Yi Liu*[⋆]*, Yunqing Xia*[⋆]*, Thomas Fang Zheng*[⋆]*, Jesper Olsen*[‡]*, and JiLei Tian*[‡]

[⋆] Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Beijing, China
[†] Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhangc@cslt.riit.tsinghua.edu.cn, {eeyliu, yqxia, fzheng}@tsinghua.edu.cn
[‡] Nokia Research Center, Beijing, China
{jesper.olsen, jilei.tian}@nokia.com

ABSTRACT

Multiple accents are often present in Mandarin speech, as most Chinese have learned Mandarin as a second language. We propose generating reliable accent specific unit together with dynamic Gaussian mixture selection for multi-accent speech recognition. Time alignment phoneme recognition is used to generate such unit and to model accent variations explicitly and accurately. Dynamic Gaussian mixture selection scheme builds a dynamical observation density for each specified frame in decoding, and leads to use Gaussian mixture component efficiently. This method increases the covering ability for a diversity of accent variations in multi-accent, and alleviates the performance degradation caused by pruned beam search without augmenting the model size. The effectiveness of this approach is evaluated on three typical Chinese accents Chuan, Yue and Wu. Our approach outperforms traditional acoustic model reconstruction approach significantly by 6.30%, 4.93% and 5.53%, respectively on Syllable Error Rate (SER) reduction, without degrading on standard speech.

Index Terms— Reliable Accent Specific Unit, Dynamic Gaussian Mixture Selection Scheme, Multiple Accents

1. INTRODUCTION

Accent variability is a significant degrading factor for most state-of-the-art automatic speech recognition (ASR). Accented speech is caused by the pronunciation differences between the speaker's first language or dialect, and that of the target speech. Such differences can be either acoustical or phonological.

There are eight major dialects in China: Guanhua, Yue, Wu, Xiang, Gan, Kejia, Minnan and Minbei, which can be further divided into more than 40 sub-categories [1]. Pronunciation difference is severe in Chinese, since written Chinese characters are ideographic and independent from their pronunciations. Seeing as most Chinese have learned Putonghua as a second language, their pronunciations are inevitably influenced by native dialects. Statistics show that over 79.58% Putonghua speakers have regional accents, and 44.03% speakers have strong accent [2]. Furthermore, multitude of accents is very often in Putonghua [3]. As a result, ASR implemented for standard Putonghua performs poorly on accented speech, especially when there are multiple accents.

Conventional methods for handling accent variations focus on modeling acoustic and phonetic variations at different levels of ASR. For phonetic variations, phone set extension and augmented pronunciation dictionary are common methods [4]. However, the extended phone set and the alternative multiple pronunciations increase lexical confusions, and do not lead to significant improvement. For acoustic variations, the most straightforward way is to build acoustic models for each accent with a large amount of accented data [5]. However, for multiple accents, an extra accent identification module is needed [6]. Another method is to apply Maximum A Posteriori (MAP) or Maximum Likelihood Linear Regression (MLLR) for acoustic adaptation to fit the acoustic characteristics of certain accents [7]. A major weakness of these approaches is the adaptations irreversibly change the parameters of the acoustic model, and make them no longer suitable for standard speech. Recently, state level pronunciation modeling and acoustic model reconstruction are applied to handle both acoustic and phonetic variations for multiple accents without sacrificing the performance on standard speech [3][8]. On the other hand, we still face challenges in the above approaches. 1) Due to the mismatch of frames, accent specific units (ASU) generated by traditional data-driven method are not reliable, which causes inaccurate modeling for accent changes. 2) Acoustic model reconstruction increases the model size and results in the low efficiency of using Gaussian mixture components in the distribution to cover accent variations.

Hence, we propose the use of reliable accent specific unit generation together with dynamic Gaussian mixture selection scheme (DGMSS) for multi-accent. The contributions are summarized as follows.

(1) Time alignment phoneme recognition is proposed to generate reliable accent specific units precisely and efficiently. Time alignment phoneme recognition is able to eliminate the mismatch of frames, which yields reliable accent specific unit candidates and their reliable training samples. Time alignment phoneme recognition is conducive to accurately model variations of each accent as well as to represent a diversity of multi-accent.

(2) In order to improve the efficiency of using Gaussian mixture components in the reconstructed distributions to handle accent variations, we propose to use dynamic Gaussian mixture selection scheme that chooses most efficient Gaussian mixtures and constructs a dynamical observation density for each speech frame in decoding progress. The selected Gaussian components increase the covering ability for accent changes and can be used efficiently in decoding. As a result, dynamic Gaussian mixture selection scheme brings preferable recognition result and relieves performance degradation

in pruned beam search without augmenting the size of acoustic models.

The paper is organized as follows. In Section 2, we present the generation of reliable accent-specific units. In Section 3, dynamic Gaussian mixture selection with acoustic model reconstruction is described. In Section 4, the recognition experiments are presented. We conclude in Section 5.

2. RELIABLE ACCENT SPECIFIC UNIT GENERATION

2.1. Chuan, Yue and Wu accents

Chuan, Yue and Wu accents were used in our paper. A speaker whose first language is a regional dialect always has a corresponding regional accent. Chuan dialect is a sub-dialect of Guanhua; Yue and Wu dialects are both major Chinese dialects. All these dialects are quite different from Putonghua in pronunciations. For example, linguists have shown only 60% of the Yue dialect pronunciations are even close to Putonghua [9]. Furthermore, differences between these dialects are also very apparent, linguists regard each of them as a distinctive language in terms of phonological, lexical and syntactic structures [9].

In Chinese ASR systems, initials and finals are usually used as subword units to construct acoustic models. Initials and finals in the three dialects are different from those in Putonghua. There are 22 initials and 36 finals for Putonghua in contrast to 20/38, 20/53 and 27/49 initials/finals for Chuan, Yue and Wu respectively [10]. The inventories of initials/finals for the four languages are distinct. For example, compared to other languages, Yue has an additional velar nasal /ng/; initials for Wu are divided into voiced and voiceless while initials for the other languages are not.

Consequently, speakers from each dialect region have difficulty in pronouncing some Putonghua initials/finals. For instance, when a Chuan dialect speaker tries to pronounce a Putonghua initial 'n', the phoneme he made may lie between 'n' and 'l', and causes an accent variation. An accent specific unit is widely used to represent an accent variation. Hence, a diversity of accent variations in multiple accent speeches is represented by different sets of accent specific units [3].

2.2. Reliable accent specific unit generation

In speech recognition, an accent variation is an erroneous recognition of a canonical phoneme into a different one due to the effect of accent. Accent specific unit $B \rightarrow S$ represents the variation that B is mis-recognized as S, where B is the canonical subword unit and S is its alternative unit. In general, accent specific unit candidates are extracted from the alignment of manually labeled canonical transcriptions and automatic generated alternative transcriptions. The alternative transcriptions are obtained by free grammar phoneme recognition [11], and contain substitution (transfer), insertion and deletion (epenthesis) errors. Since there is rare initial/final level epenthesis in Chinese accented speech, the existence of insertions and deletions causes unnecessary frame mismatch to substitutions, and results in unreliable accent specific units.

To eliminate insertion and deletion errors from alternative transcriptions, we propose time alignment phoneme recognition. Time alignment phoneme recognition decodes according to the exact duration of each phoneme, which is obtained by forced alignment [11]. Time alignment phoneme recognition performs normal recognition except for appending two additional result selection principles. 1) The number of phoneme of the selected result should be the same as that of the canonical transcription. 2) Each phoneme in the selected result should be of the same duration as its corresponding phoneme in the canonical transcription.

The procedure of generating reliable accent specific units is illustrated in Figure 1 and is explained as follows.

1) Acquire canonical transcriptions with time. To get the duration of each phoneme automatically, we perform forced alignment to phoneme-level canonical transcriptions using a pre-trained acoustic model, and get the canonical transcriptions with time.

2) Obtain alternative transcriptions. With the duration information from canonical transcriptions with time, we get the alternative transcriptions by time alignment phoneme recognition.

3) Generate reliable accent specific unit candidates. Reliable accent specific unit candidates are extracted by comparing phonemes of the same duration in canonical and alternative transcriptions.

4) Select reliable accent specific units. Reliable accent specific unit candidates include accent variations as well as errors from data and recognizer confusions [9]. Hence, we select reliable accent specific units manually from the candidates, in reference with linguistic knowledge and the confusion matrix.

The strategy for the selection in Step 4 includes the following steps: 1) Remove language inherent confusion. For instance, 'i2'→'i1' is not a typical accent variation in any of these dialects and takes high confusion probability in Putonghua. Hence, 'i2'→'i1' is language inherent confusion and is removed. 2) Replace alternative pronunciations of suspicious candidates with their inherent confusion. For example, 'an'→'ai' is not a typical Chuan change. However, Chuan speakers tend to pronounce 'an' as /æ/ that is an inex-

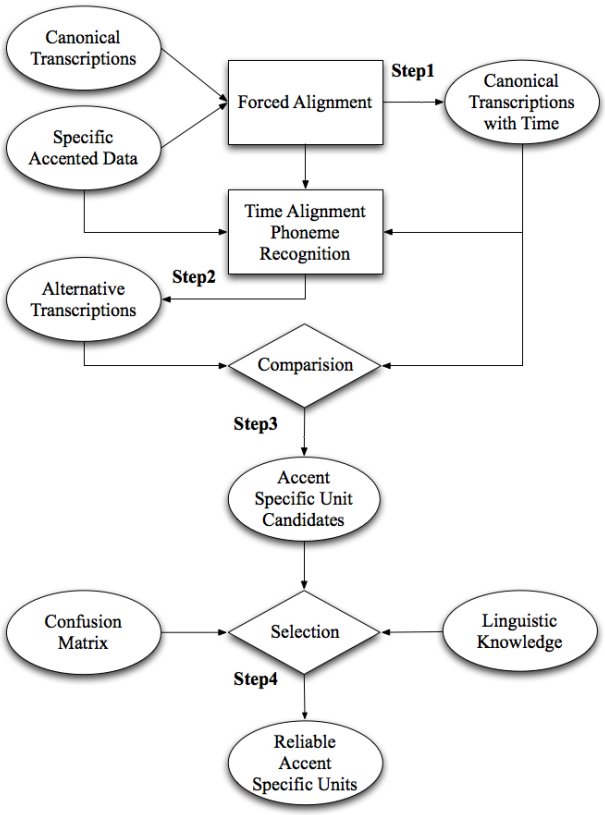

Fig. 1. Flow-chart for generating reliable accent specific units.

istent phoneme in Putonghua. Since the most similar final to /æ/ is 'ai', 'an'→'ai' is an accent variation. 3) Remove errors from either data or recognizers. For example, there is retroflex affricative 'zh' in none of the three dialects, and 'z'→'zh' does not coincide linguistic knowledge and has much less instances than 'zh'→'z'. Therefore, 'z'→'zh' is removed.

Time alignment phoneme recognition eliminates frame mismatch, and captures accurate accent variations. Compared to the traditional method using dynamic programming and edit distance [12], our method captures more accent changes (e.g. 'ch'→'s' in Chuan accent). Furthermore, time alignment phoneme recognition generates reliable instances of the candidates (e.g., 38 instances of 'un'→'en' in Chuan accent did not meet linguistic knowledge in our method while there were 61 such instances generated by traditional method), which would be used as training samples for accent specific unit models. Moreover, the selection is easier because there are always less recognition errors in time alignment phoneme recognition. Reliable accent specific units and their reliable training samples lead to accurate accent specific unit models, which play an important role in acoustic model reconstruction.

It is remarkable that although there are similar accent specific units in different accents (e.g., 'zh'→'z'), the tendency of the represented accent variations and their corresponding acoustic parameters are distinct [3]. Additionally, speakers who have lived in more than one dialectal region tend to have mixed accents. For example, in the extreme case, pronunciation of 'zh' from such speakers can distribute over the entire range between 'zh' and 'z' [3]. Therefore, this multitude of accents calls for more flexible acoustic models.

3. DYNAMIC GAUSSIAN MIXTURE SELECTION WITH ACOUSTIC MODEL RECONSTRUCTION

We build triphone acoustic models for each set of reliable accent specific units. We merge accent mixtures borrowed from such models into the pre-trained acoustic model through acoustic model reconstruction [9]. The purpose of this approach is to adjust the observation densities of the reconstructed tied-states, and increase the robustness of the models to handle accent changes along with the multitude of accents [9].

With the augmented size of the reconstructed acoustic models, we propose using dynamic Gaussian mixture selection scheme in decoding to improve the efficiency of using Gaussian mixture components in the reconstructed distributions to cover accent changes.

3.1. Acoustic model reconstruction for multiple accents

In current ASR systems, words are presented by the concatenation of subword units (e.g., phones or phonemes). The decoding formula is

$$\widehat{B} = \arg\max_{B} P(X|B)P(B), \tag{1}$$

where X is the input frame sequence, $B = b_1, b_2, \ldots, b_N$ is the canonical phoneme sequence, and N is the number of phonemes in the utterance. $P(X|B)$ is the acoustic model, and $P(B)$ is the language model that we will not consider in this paper. Due to the effect of accents, some standard subword units can be pronounced incorrectly, Equation (1) needs to be rewritten to take accent variations into consideration. Suppose $S = s_1, s_2, \ldots, s_N$ is one possible alternative pronunciation sequence, the decoding formula becomes

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

In Equation (2), $P(X|B,S)$ is the acoustic model, $P(S|B)$ is the pronunciation model. Both acoustic model $P(X|B)$ and $P(X|S)$ (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 $P(X|B,S)$ in this paper. $P(X|B,S)$ can be factorized into successive contributions

$$P(X|B,S) = \prod_{i=1}^{N} p(x_i|b_i, s_i), \tag{3}$$

where x_i is the speech frames corresponding to a canonical and alternative phoneme in the utterance. Model $p(x_i|b_i, s_i)$ is obtained by acoustic model reconstruction [13].

In this paper, decision tree based tied-state triphone model is used [11]. We build triphone models for each reliable accent specific unit. Decision trees for reliable accent specific unit models are called auxiliary trees in contrast to those for the pre-trained models are called standard trees. Since a leaf node of a decision tree represents a tied-state, acoustic model reconstruction that borrows accent mixtures from reliable accent specific unit models to augment original observation densities in pre-trained model is equivalent to merge the auxiliary tree leaf nodes into standard tree leaf nodes. An auxiliary tree leaf node is merged into a leaf node, which is nearest to it and is on the standard tree that represents the corresponding state of its canonical subword unit [9].

The new output distribution of the reconstructed tied-state $P^{'}(x|b)$ can be represented as

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

where $P(x|b)$ is the output distribution of the pre-trained model, λ is determined by the probability of the canonical phoneme be correctly recognized [9]. V is the total number of merged nodes from auxiliary trees. $P(v_i|b)$ is the confusion probability between the canonical phoneme and alternative phoneme of the accent specific unit, and can be estimated from the confusion matrix [3].

3.2. Dynamic Gaussian mixture selection scheme

In a reconstructed state, the observation density is augmented with accent mixtures to spread its coverage for handling accent variations as illustrated in part (A) and part (B) of Figure 2. Nevertheless, acoustic model reconstruction considerably increases the model size and degrades the efficiency of using Gaussian mixture components to cover accent changes. For example, in our experiment, 6,620 accent mixtures were merged into 546 standard tied-states; the state with the most accent mixtures borrowed 120 Gaussian components that belong to various accent changes and placed them at different parts of the distribution.

Dynamic Gaussian mixture selection scheme improves the efficiency of using Gaussian components to cover accent changes without further augmenting the model size. This is achieved by selecting suitable Gaussian mixtures to construct a dynamical observation density for each specified speech frame according to a k nearest mixture principle. That is, k mixtures nearest to the speech frame

are selected to customize a new output observation density for the frame. Considering the variances of Gaussian mixtures, we use Mahalanobis distance to measure the distance from a frame to a mixture. Thereby, this principle can be presented as follows.

Note $N_m = \mathbf{N}(\mu_m; \mathbf{\Sigma}_m)$, $b(\mathbf{o}) = \sum_{m=1}^{M} c_m N(\mathbf{o}; \mu_m; \mathbf{\Sigma}_m)$ is a reconstructed observation density, the Mahalanobis distance from Gaussian mixture N_m to frame $\mathbf{o}$ can be presented as

$$d_m(\mathbf{o}) = (\mathbf{o} - \mu_m)^T \mathbf{\Sigma}_m^{-1} (\mathbf{o} - \mu_m). \quad (5)$$

Suppose $N'_1, N'_2, \ldots, N'_k$ are the k mixtures nearest to $\mathbf{o}$ among all M Gaussian components, the dynamical observation density for speech frame $\mathbf{o}$ is,

$$\begin{cases} b'(\mathbf{o}) = \sum_{m=1}^{k} c''_m N(\mathbf{o}; \mu'_m; \mathbf{\Sigma}'_m) \\ c''_m = \dfrac{c'_m}{\sum_{m=1}^{k} c'_m} \end{cases} . \quad (6)$$

Moreover, k is different for different tied-states, and is determined by experiment.

For an accent frame located at the boundary of distribution, principle of the k nearest mixture selects k mixtures nearby the frame, which are the most representative Gaussian mixtures for current accent change. Gaussian mixtures that do no present the characteristics of the accent change are not included in the dynamical output distribution. Consequently, the obtained dynamical observation density has sharper borders and better model resolution ability as illustrated in part (C) of Figure 2, which lead to better covering ability for the corresponding accent change. Meanwhile, the selected Gaussian components can be used high efficiently in decoding that alleviates the performance degradation in pruned beam search. For a standard frame located at the center of the reconstructed distribution, its dynamical observation density would be similar to the original observation density before model reconstruction as shown in part (D) of Figure 2, which retains the covering ability for standard speech.

As a result, using dynamic Gaussian mixture selection scheme with acoustic models reconstructed by accent specific unit models for multi-accent, the efficiency of using Gaussian components to cover accent changes is improved. Therefore, the covering ability of the model to handle a diversity of accent variations in multi-accent is increased. Furthermore, the selected Gaussian components are most representative for a specified accent change, and can be used efficiently in decoding. As a result, dynamic Gaussian mixture selection scheme relieves the performance degradation when pruned Beam search is adopted. Meanwhile, the dynamic Gaussian mixture selection scheme neither sacrifice the performance on standard speech nor augments the model size.

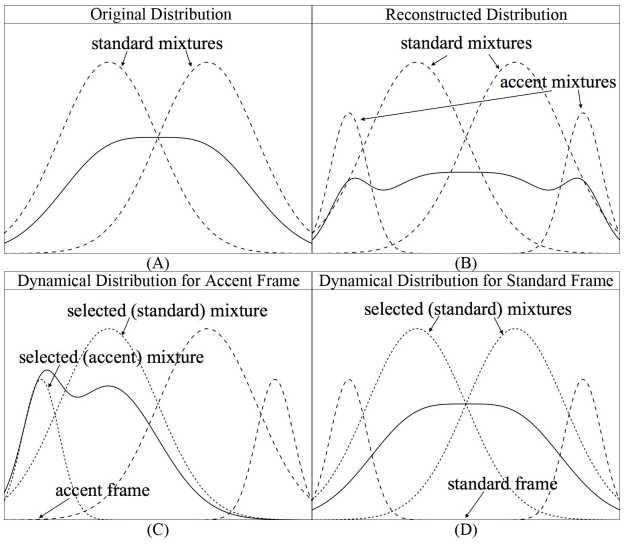

Fig. 2. Sketch map for output distributions of acoustic model reconstruction and dynamic Gaussian mixture selection scheme.

4. RECOGNITION EXPERIMENTS

We selected the development sets and testing sets for each accent from the 863 regional accent speech database [14], which is the largest and most commonly used Chinese accented speech corpus. We selected speakers with strong accents in the testing sets based on the recording records. All speech data were sampled with 16kHz and 16bit-rate, and more details are listed in Table 1. The baseline acoustic model was trained using 100 speakers' utterances with around 50 hours of Putonghua speech. It is built on HTK decision tree based state tying procedures with 3,000 tied-states triphone models and 12 Gaussian mixtures per state [11]. The HMM topology is three-states, left-to-right without skips. The acoustic features are $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. Standard Chinese 28 initials and 36 finals including 6 zero-initials were used as subword units to build HMMs. Hereafter we will use ASU to stand for accent specific unit, and use DGMSS to take the place of dynamic Gaussian mixture selection scheme for short, in all tables and figures.

165, 191 and 166 reliable accent specific units were generated from DevC, DevY and DevW respectively. We constructed 495 auxiliary trees with 517 tied-states for Chuan, 573 auxiliary trees with 605 tied-states for Yue and 498 auxiliary trees with 533 tied-states for Wu. These tied-states from accent specific unit models were merged into the baseline AM through acoustic model reconstruction. The reconstructed acoustic model (System 3) included 42,620 mixtures with 14.2 mixtures per state on average. In order to show the effectiveness of reliable accent specific units, 160, 187 and 166 traditional accent specific units were extracted from the alignment on DevC, DevY and DevW individually, which were generated by Flexible Alignment Tool [12]. Then, 480, 561 and 498 auxiliary trees with 531, 582 and 569 tied-states were built for Chuan, Yue and Wu unit correspondingly. The reconstructed acoustic model included 42,728 mixtures (System 2). Moreover, we got System 4 by using dynamic Gaussian mixture selection scheme with the acoustic models from System 3. Different k (number of selected mixtures) for each tied-state are determined using development sets. In our experiment, k ranges from 1 to the number of mixtures in the state.

Table 2 shows the effectiveness of our proposed approach evaluated in free grammar Chinese syllable recognition task. Compared to the Baseline (System 1), acoustic model reconstruction with traditional accent specific unit (System 2) yields significant SER reduction on every accented testing set. The reason lies in the fact that accent mixtures in the reconstructed tied-states adjust the original distribution and enable more Gaussians at the boundaries to cover confusing pronunciation of accent changes [3]. Compare to System 2, System 3 gives 2.26%, 1.12% and 1.89% relative SER reduction on TestC, TestY and TestW respectively. These results indicate that the reliable accent specific units have better covering ability for accent changes than the traditional accent specific units.

With dynamic Gaussian mixture selection scheme, System 4 obtains 4.13%, 3.85% and 3.71% lower relative SER than System 3. The selected efficient Gaussian components increased the covering ability for variations of multi-accent. Thereby, the joint use of re-

Table 1. Data sets separation in experiments.

ID	DevC	TestC	DevY	TestY	DevW	TestW	TestP
Duration	6.5h	2.2h	6.1h	1.9h	6.6h	2.1h	1.8h
Syllable Number	51,907	18,824	51,341	18,363	52,584	17,666	9,055
Speaker Number	20	20	20	20	20	20	10
Utterance Number	3,205	1,000	3,091	1,000	3,471	1,000	1,000
Speech Type	Chuan accent		Yue accent		Wu accent		Putonghua

Table 2. Lower SER for using our approach compared to using traditional accent specific unit and MAP adaptation.

ID	System	Syllable Error Rate (SER) %			
		TestP	TestC	TestY	TestW
1	**Baseline**	23.45	52.02	53.36	54.05
2	**Reconstructed HMMs (Traditional ASU)**	22.92 (-0.53)	43.78 (-8.24)	44.63 (-8.73)	43.94 (-10.11)
3	**Reconstructed HMMs (Reliable ASU)**	22.98 (-0.47)	42.79 (-9.23)	44.13 (-9.23)	43.11 (-10.94)
4	**Reconstructed HMMs (Reliable ASU) + DGMSS**	22.97 (-0.48)	41.02 (-11.00)	42.43 (-10.93)	41.51 (-12.54)
5	**MAP adaptation with DevC, DevY and DevW**	32.71 (+9.26)	43.91 (-8.11)	44.63 (-8.73)	43.41 (-10.64)

liable accent specific units and dynamic Gaussian mixture selection scheme (System 4) achieves 6.30%, 4.93% and 5.53% lower relative SER than the traditional acoustic model reconstruction approach (System 2) for TestC, TestY and TestW respectively.

Comparing System 4 to System 5, System 4 achieves significantly 6.58%, 4.93% and 4.38% lower relative SER reduction than System 5. These results show that explicitly and accurately modeling each accent change is better than adapting a model to fit accent changes for all accents together. Meanwhile, our approach does not degrade on standard speech, while MAP adaptation severely does. MAP adaptation adjusts the parameter of acoustic models to fit multi-accent, and makes them no longer suitable for standard speech.

An example for using dynamic Gaussian mixture selection scheme to reduce local model mismatch for Yue accent is presented in Figure 3. In this example, when using Baseline and System 3, final 'ing' of syllable 'jing' was mis-recognized as 'in'. This was caused by an accent variation between 'ing' and 'in' in Yue accent. The 3 states of the mis-recognized final are presented from frame 208 to 217. The acoustic scores of both Baseline and acoustic model reconstruction with reliable accent specific unit severely dropped around frame 209. In the system of acoustic model reconstruction with reliable accent specific unit, the borrowed accent mixtures 'ing'→'in' helped to increase the acoustic score around frame 211, but not enough to restore the local mode mismatch. However, with dynamic Gaussian mixture selection scheme, representative Gaussian mixtures were selected for the accent samples. The generated dynamical observation densities further improved the covering ability for 'ing'→'in'. Therefore, System 4 successfully restored this local model mismatch and gave a correct recognition result.

Finally, we will show the benefit that dynamic Gaussian mixture selection scheme brings to pruned beam search. Table 3 presents the performance under different pruning degrees. Parameter t means any model whose maximum acoustic score falls more than the value of t below the maximum among all of the models will be deactivated [11]. Relative SER reductions to the best (i.e., no-pruned, $t = 0$) results are presented in brackets. From Table 3, we can see pruning significantly reduced the time cost for decoding at the price of decreasing the recognition accuracy in different degrees. In addition, dynamic Gaussian mixture selection scheme considerably relieved the performance degradation in all cases. The reason lies that dynamic Gaussian mixture selection scheme chooses the most efficient Gaussian mixtures that not only handle accent changes but can be used with high efficiency in decoding as well.

Therefore, dynamic Gaussian mixture selection improves the presentation of acoustic model reconstruction in pruned beam search, which is necessary in a real ASR system. Moreover, it is remarkable that the benefits from dynamic Gaussian mixture selection scheme averagely costs about 24% more time. This additional time consuming is not only spent on the execution of dynamic Gaussian mixture selection itself, but also on expanding different searching paths. The execution time of dynamic Gaussian mixture selection can be reduced by a better implementation of the algorithm.

5. CONCLUSIONS

We have presented using time alignment phoneme recognition to efficiently generate reliable accent specific units, which capture variations accurately for multi-accent, and such units are able to model

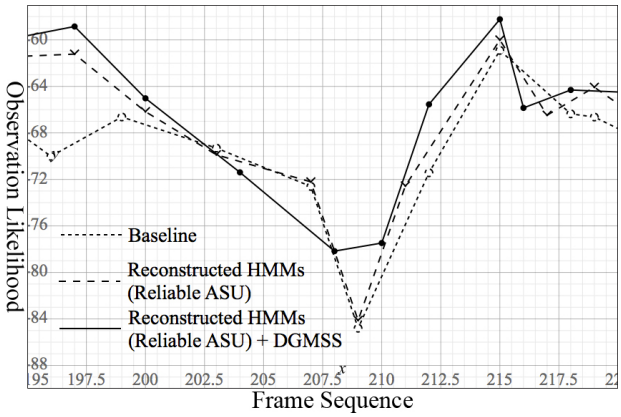

Fig. 3. Use dynamic Gaussian mixture selection scheme to restore local model mismatch in decoding.

Table 3. Performance degradation in pruned beam search.

Pruning threshold	Reconstructed HMMs (Reliable ASU)		Reconstructed HMMs (Reliable ASU) + DGMSS	
	SER%	Time(s)	SER%	Time(s)
$t = 0$	44.13	37,093.67	42.43	46,992.71
$t = 250$	44.59 (+1.04)	29,751.38	42.48 (+0.12)	38,465.84
$t = 200$	45.41 (+2.90)	20,031.06	42.95 (+1.23)	25,041.35
$t = 150$	47.73 (+8.16)	10,132.87	45.29 (+6.74)	12,105.82
$t = 100$	68.05 (+54.20)	2,447.08	57.98 (+36.65)	2,932.51

accent changes explicitly at both phonetic and acoustic level. Meanwhile, we propose dynamic Gaussian mixture selection scheme to generate a dynamical observation density for each speech frame by selecting the most efficient Gaussian mixture components. As a result, the approach of dynamic Gaussian mixture selection improves the covering ability for a diversity of accent changes in multi-accent and alleviates the performance degradation caused by pruned beam search without augmenting the model size. Experimental results show the joint use of these two methods yields 6.30%, 4.93% and 5.53% relative SER reduction on Chuan, Yue and Wu accents individually, compared to previous acoustic model reconstruction with traditional accent specific unit. Compared with MAP adaptation, our approach achieves 6.58%, 4.93% and 4.38% SER reduction on each certain accent respectively, without sacrificing the performance on standard Putonghua.

In future work, we plan to investigate an alternative way to select reliable accent specific units from the candidates automatically. In addition, other selection principles for dynamic Gaussian mixture selection scheme and what if some discriminative training is jointly used with our methods will also be investigated.

6. ACKNOWLEDGEMENT

This work was support by Natural Science Foundation of China (60975018), the joint research grant of Nokia-Tsinghua Joint Funding 2008-2010, and New Teacher Grant of Ministry of Education of China (20090002120012).

7. REFERENCES

[1] J. Li, T.-F. Zheng, W. Byrne, and D. Jurafsky, "A dialectal Chinese speech recognition framework," *Journal of Computer Science and Technology*, vol. 21, pp. 106–115, 2006.

[2] Leading Group Office of Survey of Language Use in China, *Survey of Language Use in China (in Chinese)*, Yu Wen Press, Beijing, 2006.

[3] Y. Liu and P. Fung, "Multi-accent Chinese speech recognition," in *Interspeech*, Pittsburg, U.S. Pennsylvania, September 2006, pp. paper 1887–Mon1BuP.8.

[4] G.-H. Ding, "Phonetic confusion analysis and robust phone set generation for Shanghai-accented Mandarin speech recognition," in *Interspeech*, Brisbane, Australia, September 2008, pp. 1129–1132.

[5] V. Fisher, Y. Gao, and E. Janke, "Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognition," in *International Conference on Spoken Language Processing*, Sydney, Australia, November 1998, p. paper 0223.

[6] F.S. Richardson, W.M. Campbell, and P.A. Torres-Carraquillo, "Discriminative n-gram selection for dialect recognition," in *Interspeech*, Brighton, U.K., September 2009, pp. 192–195.

[7] Y.R. Oh and H.K. Kim, "MLLR/MAP adaptation using pronunciation variation for non-native speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, November 2009.

[8] M. Saraclar, H. Nock, and S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, pp. 137–160, 1999.

[9] Y. Liu and P. Fung, "Partial change accent models for accented Mandarin speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, December 2003.

[10] J.-H. Yuan et.al., *Survey of The Chinese dialects (in Chinese)*, Yu Wen Press, Beijing, 2nd edition, 2001.

[11] S. Young et.al., *The HTK Book*, Entropic Cambridge Research Laboratory, 3.4 edition, 2009.

[12] P. Fung, W. Byrne, T.-F. Zheng, T. Kamm, Y. Liu, Z.-J. Song, V. Venkataramani, and U. Ruhi, "Pronunciation modeling of Mandarin casual speech," Tech. Rep., Summer Research Workshop, John Hopkins University, 2000.

[13] Y. Liu and P. Fung, "Pronunciation modeling for spontaneous Mandarin speech recognition," *International Journal of Speech Technology*, vol. 7, pp. 155–172, 2004.

[14] ChineseLDC.org, *Resource Introduction to RASC863*, <www.chineseldc.org/doc/CLDC-SPC-2004-005/intro.htm>, 2009.