Asymmetric Acoustic Model for Accented Speech Recognition

Chao Zhang[*†], Yi Liu[*], Thomas Fang Zheng[*]
[*]Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Beijing, China
E-mail: zhangc@cslt.riit.tsinghua.edu.cn, {eeyliu, fzheng}@tsinghua.edu.cn Tel: +86-10-62796589
[†]Department of Computer Science and Technology, Tsinghua University, Beijing, China

***Abstract*—We propose to improve accented speech recognition performance by using asymmetric acoustic model. Our proposed model is generated based on reliable accent specific units and acoustic model reconstruction. The reliable units are extracted with time alignment recognition to cover accent variations at both acoustic and phonetic levels. The asymmetric acoustic model is obtained through selective decision tree merging together with dynamic Gaussian component selection in model reconstruction. The improved resolution of our proposed model is able to handle different levels of accented variations at different degrees. The effectiveness of our approach was evaluated on a typical Chinese accent. Our system outperforms traditional acoustic model reconstruction and MAP adaptation approaches by 8.28% and 7.14%, relatively on Syllable Error Rate (SER) reduction without sacrificing the performance on standard Mandarin speech.**

I. INTRODUCTION

Most state-of-the-art automatic speech recognition (ASR) systems fail to perform well when the speaker has a regional accent. Accent is a serious problem for Chinese speakers since most of Chinese learn standard Mandarin (Putonghua) as a second language, and their pronunciations are strongly influenced by the native regional dialects [1]. There are eight major dialectal regions in China, which can be further divided into more than 30 sub-categories. The acoustic and linguistic representations are quite different between Putonghua and the dialects. Therefore, the pronunciation of Putonghua is inevitably influenced by the native dialect of the speaker. Statistics show that over 79.58% Putonghua speakers have regional accents, and 44.03% speakers have strong accent [2]. Accented speech differs from standard speech in terms of phonological, morphological, syntactic and lexical characteristics. As a result, ASR systems implemented for processing standard Mandarin usually perform poorly for non-native accented speech.

Modeling accent effects at acoustic and phonetic levels are commonly used in recently researches [3][4][5][6][7]. Maximum A Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR) adaptation and using discriminative training to refine acoustic models are efficient and straightforward approaches to handle accents at acoustic level [3][6]. Phone set extension and augmenting the pronunciation dictionary with accent specific units and their relevant probabilities are two typical methods at phonetic level in modeling accent variations [4][7]. A major weakness in above approaches is that MAP or MLLR irreversibly changes parameters of acoustic models, which makes them no longer suitable for multiple accents as well as native standard speech recognition. Meanwhile, the extended phone set and augmented pronunciations in dictionary may introduce more lexical confusion in decoder. The pronunciation changes in different accents are very flexible, and cannot be accurately modeled by using alternative phones only [5]. State level pronunciation modeling and acoustic model reconstruction was proposed to handle both acoustic and phonetic variations without degrading on standard speech [5][8]. However, we still face challenges in these approaches: the generated accent specific units (ASU) are not reliable due to the mismatch of frames caused by traditional data-driven method; the asymmetry of accent changes is not fully exploited.

In this paper, we propose to use asymmetric acoustic model for accented speech recognition, and such model is generated based on reliable accent specific units and acoustic model reconstruction. To get reliable accent specific unit, we apply time alignment recognition that aims to eliminate frame mismatch by recognition according to accurate phoneme boundary information. Through the use of time alignment recognition, we are able to generate and select accent specific unit accurately and efficiently. We obtain their dependable training samples, and acquire better modeling for accent changes at both phonetic and acoustic levels. Subsequently, together with selective decision tree merging for acoustic model reconstruction, we use dynamic Gaussian selection to utilize the asymmetry of accent changes. Dynamic Gaussian selection selects appropriate Gaussian components to build a dynamical observation density for each specified speech frame in decoding. Therefore, the generated asymmetric acoustic model improves acoustic model resolution and handles different levels of accent variations at different degrees. Experiments show that our proposed method yields better recognition results than traditional acoustic model reconstruction method and MAP adaption, in addition to not degrading on standard speech.

In Section 2, we introduce using time alignment recognition to generate reliable accent specific unit. In Section 3, we describe our approach of asymmetric acoustic modeling using model reconstruction and dynamic Gaussian selection. Section 4 and Section 5 are experimental results and

conclusions.

II. Reliable Accent Specific Unit Generation

In speech recognition, an accent change is an erroneous recognition of a canonical phoneme into a different one caused by the accented pronunciation variation made by the speaker. An accent specific unit is commonly used to represent an accent change, and is typically noted as $B \rightarrow S$ in which B is the canonical phoneme and S is its alternative pronunciation [5]. It is remarkable that this asymmetric notation is in conformity with the asymmetric characteristic of accent confusions [9].

A common approach for obtaining alternative phoneme sequence is through free grammar recognition [1]. In general, such obtained sequence usually contains a large amount of insertion and deletion errors, which cause frame mismatch and result in unreliable accent specific units. For the sake of eliminating frame mismatch, we implement time alignment recognition to get the alternative phonemes with no insertions and deletions. Time alignment recognition acquires the phoneme boundary information in each utterance by forced alignment [10], and performs normal recognition except for appending an additional principle to select the result: the selected result should have the same number of phonemes as its canonical transcription, in the restriction that the duration for each alternative phoneme should coincide with its corresponding canonical phoneme boundary.

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

1) Obtain canonical transcriptions with phoneme boundary information. We perform forced alignment to phoneme-level canonical transcriptions using pre-trained acoustic models to get the duration information for each phoneme.

2) Produce alternative transcriptions. With the duration information from the transcriptions generated in step 1), we use time alignment recognition to generate the alternative transcriptions.

3) Generate accent specific unit candidates. We extract the candidates by comparing phonemes at corresponding positions in canonical and alternative transcriptions.

4) Select accent specific units. Accent specific unit candidates contain accent changes as well as errors from data and recognizer confusions [1]. Thus, we select reliable accent specific units manually from the candidates, in reference with linguistic knowledge and confusion matrix.

The strategy for filtering accent specific unit candidates in step 4) includes following parts. 1) Remove errors from data and recognizer (e.g. 'n' $\rightarrow$ 'd'). 2) Remove language inherent confusions (e.g. 'i2' $\rightarrow$ 'i1'). 3) Replace alternative pronunciation of a suspicious candidate with its inherent confusion. For example, Sichuan accent speakers tend to pronounce 'an' as /ae/ that is an inexistent pronunciation in Putonghua. Since 'ai' is the most similar pronunciation to /ae/ in Putonghua, 'an' $\rightarrow$ 'ai' is selected. Both errors from data or recognizer and language inherent confusions are obtained by linguistic rules and generated confusion matrix [5].

Time alignment recognition avoids frame mismatch and is able to capture accurate accent changes. Compared to previous method using free grammar recognition and Flexible Alignment Tool with cost transducer [1], the proposed method can cover a diversity of flexible accent changes (e.g. 'eng' $\rightarrow$ 'uan'). Furthermore, our proposed method generates reliable instances for the candidates. For example, 29 instances of 'n' $\rightarrow$ 'l' do not meet linguistic knowledge produced by our method in contrast to 77 such instances generated by the traditional method. Moreover, the selection is easier since there are fewer errors in time alignment recognition.

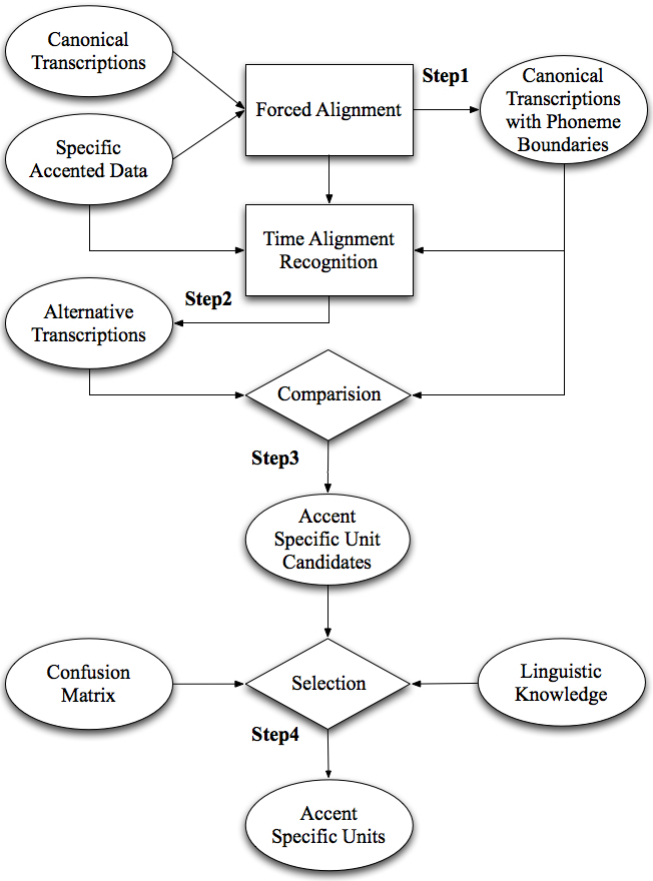

Fig. 1 Flow-chart for reliable generation procedure of accent specific units.

III. Asymmetric Acoustic Modeling

To build asymmetric acoustic model, we first perform acoustic model reconstruction with reliable accent specific units, then integrate dynamic Gaussian selection. For acoustic model reconstruction, we build triphone acoustic models for each reliable accent specific unit, and merge Gaussian components from accent specific unit models into the pre-trained triphone acoustic models through decision tree merging, improving the robustness ability of the pre-trained models to cover various accent changes. Afterwards, our proposed approach of dynamic Gaussian selection is adopted to further utilize the asymmetry of accent variations by

selecting suitable Gaussian components for each specified speech frame, and results in the asymmetric acoustic model.

A. Acoustic Model Reconstruction

Decision tree based tied-state triphone model is used in our work [10]. Decision trees for accent specific units are called auxiliary trees in contrast to those for the pre-trained acoustic models are called conventional trees. Since every leaf node of a decision tree represents a tied-state, borrowing Gaussian components from an accent specific unit model into a pre-trained model, that makes the pre-trained model be able to cover accent changes, equals to merge an auxiliary tree leaf node into a standard tree leaf node, as illustrated in Fig. 2.

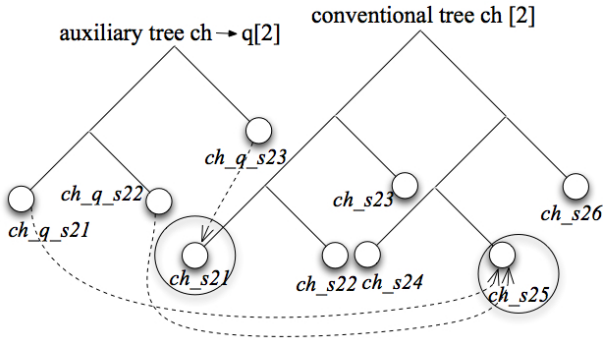

Fig. 2 Sketch map for decision tree merging.

Moreover, the mapping relationship between an auxiliary tree leaf node and a standard tree leaf node is decided by minimum distance, which is measured using the asymmetric distance in accordance with the asymmetry of accent changes [9]. More details about acoustic model reconstruction can be found in previous work [5].

B. Dynamic Gaussian Selection

For a reconstructed tied-state, the observation density is enlarged with accent mixtures from reliable accent specific unit models to handle accent changes, as illustrated in part (A) and (B) in Fig. 3. To further improve modeling accent changes, we use dynamic Gaussian selection. This approach builds a dynamic observation density for each speech frame by selecting suitable mixtures from the reconstructed state according to a k-nearest mixtures principle, namely we select k mixtures nearest to the specified frame. In addition, to better meet the directional asymmetry nature of Gaussian distribution (i.e., the variance can be different in different dimensions), we choose the Mahalanobis distance to measure the distance from a frame to a Gaussian component, which is obviously an asymmetric distance in substance.

Dynamic Gaussian selection is explained as follows.

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

$$d_m(\mathbf{o}) = (\mathbf{o} - \mu_m)^{\mathrm{T}} \Sigma_m^{-1} (\mathbf{o} - \mu_m). \tag{1}$$

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

$$\begin{cases} b'(\mathbf{o}) = \sum_{m=1}^{k} c''_m \mathrm{N}(\mathbf{o}; \mu'_m; \Sigma'_m) \\ c''_m = c'_m \Big/ \sum_{m=1}^{k} c'_m \end{cases}. \tag{2}$$

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

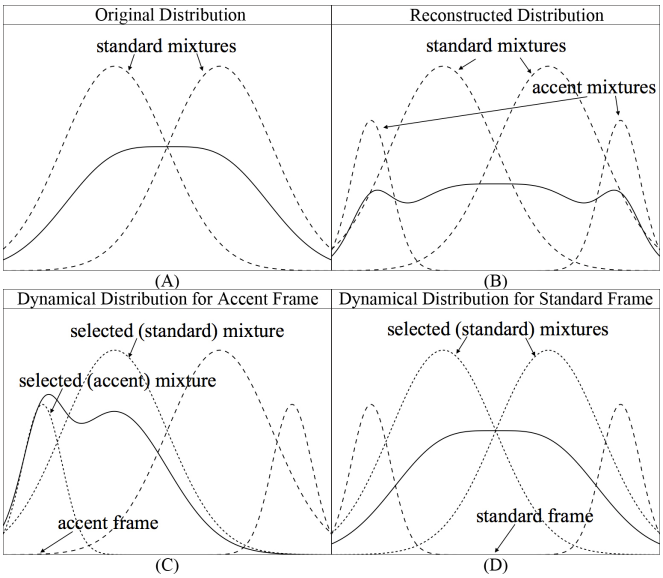

Fig. 3 Sketch map for distributions of acoustic model reconstruction and dynamic Gaussian selection.

Integrating dynamic Gaussian selection, the Gaussian distribution will be adjusted according to the parameter k, namely the selected Gaussian component numbers. For an accented frame located at the boundary of distribution, k mixtures nearby the boundary will be selected, and the achieved dynamical observation density has sharper borders as illustrated in part (C) of Fig. 3. With appropriate k, dynamic Gaussian selection is able to model different levels of accent changes at different degrees. Furthermore, it is well known that accent variations are asymmetric, that is, a pronunciation variation might be completely different from its inverse variation (e.g. 'zh'→'z' and 'z'→'zh') in terms of acoustic and phonetic parameters [9]. Therefore, using dynamic Gaussian selection better fits this asymmetry of accent changes, and improves the model resolution ability.

Moreover, for a standard speech frame located at the center of distribution, the dynamical observation density for it would be similar to the original distribution as shown in part (D) of Fig. 3. As a result, dynamic Gaussian selection retains the covering ability for standard speech.

IV. EXPERIMENTS

We evaluated our approach in a Chinese desktop sentence speech recognition task. There is no word n-gram in these sentences so that we can isolate the effect of our approach without the influence from high-level information.

The database includes both Putonghua and a typical Chinese accent-Sichuan for comparable experiment. The database was male and female balanced. All speech data were sampled with 16kHz and 16bit-rate. A detailed data set separation in our experiments is described in Table 1. PTH and SC stand for Putonghua and Sichuan accent, respectively.

TABLE I
THE SEPARATION OF DATASET IN EXPERIMENTS

ID	Train-PTH	Test-PTH	Dev-SC	Test-SC
Speech Type	Putonghua		Sichuan accent	
Duration	51.5h	1.8h	5.3h	2.9h
Syllable Number	340,556	16,214	51,094	29,870
Speaker Number	100	10	26	10
Utterance Number	25,920	1,000	1,734	1,000

The HMM topology is three-states, 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 were used to generate context-independent HMMs. We built 12 Gaussian mixtures triphone models with 3,000 tied-states using HTK decision tree based state tying procedures [10]. Dictionary with 413 syllables was used in all experiments.

84 reliable ASUs were generated from Dev-SC. We constructed 252 auxiliary trees with 257 tied-states. Through acoustic model reconstruction, tied-states from reliable ASU models were merged into the pre-trained acoustic models with 3,000 tied-state and 192 conventional trees. The reconstructed acoustic models contained 37,028 Gaussian mixtures. In order to show the effectiveness of reliable ASU, a control group was reconstructed with 89 traditional ASUs from Dev-SC. We constructed 267 auxiliary trees with 271 tied-states. The reconstructed acoustic model had 37,084 mixtures.

TABLE II
LOWER SER FOR USING OUR APPROACHES COMPARED TO USING TRADITIONAL ASU AND MAP ADAPTATION

	System	Syllable Error Rate (SER)%	
		Test-SC	Test-PTH
1	Baseline	41.00	21.70
2	Reconstruct HMMs with Traditional ASU	35.85 (-5.15)	21.49 (-0.21)
3	Reconstruct HMMs with Reliable ASU	34.51 (-6.49)	21.55 (-0.15)
4	Reconstruct HMMs with Reliable ASU + Dynamic Gaussian Selection	32.88 (-8.12)	21.52 (-0.18)
5	Baseline + MAP Adaptation with Dev-SC	35.41 (-5.59)	29.96 (+8.26)

Table 2 shows that accent gives an inverse impact on recognition accuracy when acoustic model is trained only from standard speech. k for different tied-states are determined using Dev-SC. Compared to Baseline, the reconstructed HMMs with traditional ASU yield a significant 5.15% absolute SER reduction on Test-SC. This result indicates that accent mixtures in reconstructed states adjusts the original mixture distribution and enables more Gaussians at boundaries to cover accent changes [1].

It is shown in Table 2 that System 3 gives 1.34% absolute SER reduction with respect to System 2. This result proves the reliable ASU models are more accurate than traditional ASU models, and are able to cover more accent changes. With the integration of dynamic Gaussian selection to System 3, an additional 1.63% absolute SER reduction is achieved on accented speech. It is shown that asymmetric acoustic model improves resolution ability for accent changes by covering variations at different levels with different degrees and better fits the asymmetric characteristic of accent changes. As a result, the joint use of our proposed methods (System 4) yields a significant 8.28% relative SER reduction than traditional acoustic model reconstruction (System 2).

Comparing System 4 to System 5, the proposed approach outperforms MAP adaptation significantly by 7.14% relative SER reduction, and does not degrade on standard speech while MAP adaptation severely does. The reason lies in the fact that the MAP adaptation irreversibly changes the parameters of acoustic models that make them no longer appropriate for standard speech.

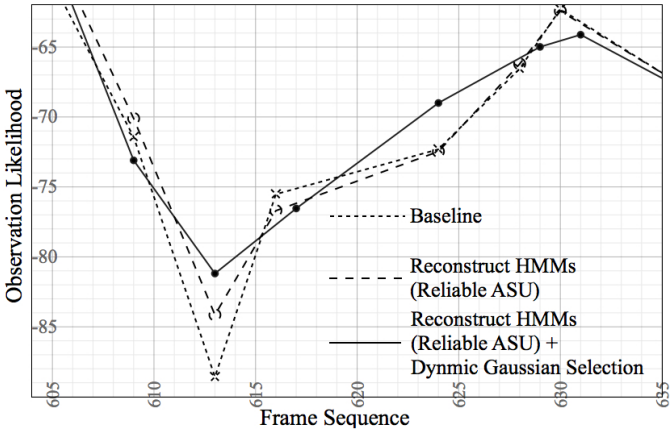

Fig. 4 Using asymmetric acoustic model to restore local model mismatch in decoding.

An example for using asymmetric acoustic model to reduce local model mismatch is presented in Fig. 4. When using Baseline and System 3, initial 'ch' was misrecognized as 'c'. This was caused by a pronunciation variation from 'ch' to 'c' in Sichuan accent, and the acoustic score drops significantly between frame 610 to frame 615. In System 3, accent mixtures from reliable ASU model 'ch'→'c' can increase the acoustic score but not enough to obtain high acoustic score to output correct recognition result. However, in System 4, as dynamical Gaussian selection selects appropriate mixtures for accented frames, the dynamic observation densities strengthen the effect of accent mixtures relatively from 'ch'→'c', thus

successfully restores this local model mismatch and gives a correct recognition result.

V. CONCLUSIONS

We have described asymmetric acoustic model for accented speech recognition. This model is built on model reconstruction with reliable accent specific units, as well as dynamic Gaussian selection. Time alignment recognition is used to capture a diversity of accent changes according to phoneme boundaries. Dynamic Gaussian selection selects suitable Gaussian components to construct a dynamic observation density for each speech frame according to k-nearest mixture principle. Asymmetric acoustic model handles accent variations of different levels at different degrees, and retains the covering ability for standard speech. Experimental results showed asymmetric acoustic model yielded 8.28% relative SER reduction than traditionally acoustic model reconstruction, and achieved 7.14% SER reduction compared to MAP adaptation without degrading on standard speech.

ACKNOWLEDGMENT

We would like to thank Prof. Chinhui Lee of School of Electrical and Computer Engineering, Georgia Institute of Technology for his valuable and instructive suggestion as well as providing useful tools in this paper. This work was support by Natural Science Foundation of China (60975018), the joint research grant of Nokia-Tsinghua Joint Funding 2008-2010.

REFERENCES

[1] Y. Liu and P. Fung, "Partial change accent models for accented Mandarin speech recognition," in *Proc. of the IEEE ASRU*, 2003.

[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.R. Oh and H.K. Kim, "MLLR/MAP adaptation using pronunciation variation for non-native speech recognition," in *Proc. of the IEEE ASRU*, 2009.

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

[5] P. Fung and Y. Liu, "Effects and modeling of phonetic and acoustic confusions in accented speech recognition," *Journal of the Acoustical Society of America*, Vol.118, Issue 5, pp.3279 3293, Nov. 2005.

[6] D. Vergyri, L. Lamel, J.L. Gauvain, "Automatic speech recognition of multiple accented English data," in *INTERSPEECH*, 2010, 1652-1655.

[7] L.-Q. Liu, F. Zheng, et.al., "Using a small development data set to build a robust dialectal Chinese speech recognizer," in *INTERSPEECH*, 2007, 1729-1732.

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

[9] M.-Y. Tsai and L.-S. Lee, "Pronunciation variation analysis based on acoustic and phonetic distance measure with application examples on Mandarin Chinese," in *Proc. of the IEEE ASRU*, 2003.

[10] S. Young et.al., *The HTK book*, Entropic Cambridge Research Laboratory, 2009.