Multi-Accent Chinese Speech Recognition

LIU Yi and Pascale Fung

Human Language Technology Center
Department of Electrical and Electronic Engineering
and Velda Limited
University of Science and Technology, Clear Water Bay, Hong Kong
{eeyliu, pascale@ee.ust.hk}

10.21437/Interspeech.2006-34

Abstract

Multiple accents are often present in spontaneous Chinese Mandarin speech as most Chinese have learned Mandarin as a second language. We propose a method to handle multiple accents as well as standard speech in a speaker-independent system by merging auxiliary accent decision trees with standard trees and reconstruct the acoustic model. In our proposed method, tree structures and shape are modified according to accent-specific data while the parameter set of the baseline model remains the same. The effectiveness of this approach is evaluated on Cantonese and Wu accented, as well as standard Mandarin speech. Our method yields a significant 4.4% and 3.3% absolute word error rate reduction without sacrificing the performance on standard Mandarin speech.

Index Terms: multi-accent recognition, accent tree, Chinese.

1. Introduction

Automatic recognition of speaker-independent, natural spontaneous speech often needs to handle accents [1]. Accented speech is caused by the difference in pronunciation between the speaker's first language or dialect source, and that of target speech. Such difference can be acoustical or phonological.

Most speakers of Mandarin Chinese learned Mandarin (Putonghua) as second language, and their pronunciations are strongly influenced by their native regional language. Very often, there is a multitude of accents present in the pronunciation of Mandarin speech by Chinese speakers [2, 3]. As a result, ASR systems implemented for processing standard Putonghua, perform poorly for non-native accented speech, especially when there are multiple accents

Conventional methods to handle accented speech are to focus on modeling phonetic and acoustic changes [2, 3, 4, 5, 6]. Phone set extension to include accent-specific units is a common way to model phonetic changes. However, the extended phone set and augmented pronunciations may introduce more lexical confusion in the decoder. Acoustic model parameters are commonly modified to model acoustic changes in accented speech. This type of approach includes retraining acoustic model using a large amount of accented speech [2]; applying Maximum A Posteriori (MAP) or Maximum Log Likelihood Ratio (MLLR) adaptation to fit the characteristics of a particular accent [5, 6]; and using discriminative training to refine acoustic models [7]. A major weakness in these approaches is that the parameters of acoustic models undergo an irreversible change, and the models lose their ability to cover other accents.

In this paper, we propose a method to handle multiple regional accents together with standard speech, in a speaker-independent scenario. We extend on our previous work on accented Mandarin speech recognition [3], by using different sets of accent-specific units with acoustic model reconstruction. A diversity of accent changes in multiple accented speech are represented by different sets of accent-specific units. The acoustic model reconstruction is performed at state level through decision tree merge on state-tying triphone ASR system. We focus on two specific major accents in Chinese speakers – Cantonese and Wu accents. All speakers are fluent in Mandarin while their first language is either Cantonese or Wu.

2. Multiple Accents in Chinese Speech

In addition to standard Mandarin (also known as Putonghua) spoken by radio and TV announcers, there are seven major language regions in China including Guanhua (Mandarin), Wu, Yue (Cantonese), Xiang, Gan, Min and Kejia [8]. These major languages can be further divided into more than 30 sub-categories of dialects.

Table 1 lists the accent distribution of 2412 speakers in the HKUST Mandarin Telephone Speech corpus [10]. In general, there are seven Sinitic language groups in China. The second column illustrates general accent distribution of Mandarin speakers [8]. The third column gives the distribution in the HKUST corpus. "Unknown" means that the speaker did not provide any accent information.

Accent regions	Distribution in general public	Distribution in the corpus
Guanhua	70%	77.1%
Wu	8.4%	8.8%
Cantonese	6%	8.4%
Xiang	5%	2.2%
Gan	2.4%	1%
Min	4.2%	1.7%
Kejia	4%	0.3%
Unknown	--	0.45%

Table 1: Mandarin accent distribution.

In addition to lexical, syntactic and colloquial differences, the phonetic pronunciations of the same Chinese characters are quite different between Putonghua and other Sinitic languages. Linguists have shown, for example, that only 60% of Cantonese

and 70% of Wu pronunciations are even *close* to Putonghua pronunciations [8]. Therefore, although Putonghua is widely spoken in China, its pronunciation is still strongly influenced by the native language of the speaker. Speakers who have lived in several geographical regions also tend to have mixed accents. This type of accent variations is obviously beyond straight forward phone changes.

In Chinese ASR systems, initial and final units are commonly used as subword units instead of phonemic units [2, 3, 4]. There are 21 initials and 37 finals for Putonghua in contrast to 19 initials and 53 finals in Cantonese and 35 initials and 32 finals in Wu dialect [8]. All initials consist of a single consonant. However, the initial inventories for Putonghua and two other languages are distinct. For example, in contrast to Putonghua initials, Cantonese initials do not have retroflexed affricatives (e.g., 'zh, ch, sh, r'), and has one additional velar nasal 'ng'. The structure of Cantonese finals is more complex than that of Putonghua. Cantonese finals have six different consonant codas ('m', 'n', 'ng', 'k', 'p' and 't') in contrast to the two codas 'n' and 'ng' in Putonghua finals. Wu dialect initials do not have retroflexed affricatives either. The pronunciations of 'zh, ch, sh' are moved to 'z, c, s', but the pronunciations are not completely the same as the Putonghua 'z, c, s's. In addition, voiced initials in Wu include more phonetic changes compared to Putonghua. For finals, there is no pronunciation difference between 'n' and 'ng', and the number of monophthong is larger than that of Putonghua (e.g., the monophthongs 'ai' and 'ei' in Putonghua are changed to 'a' and 'e' in Wu).

Consequently, Cantonese and Wu speakers often have difficulty pronouncing some basic Putonghua initials/finals, which leads to pronunciation changes. The standard Putonghua initial set cannot represent the actual pronunciations in accented speech. In an extreme case, if the speaker's pronunciation is affected by both Cantonese and Wu dialects, the pronunciation of 'zh' can distributed over the entire range between 'zh' and 'z'. Hence, a more powerful acoustic model is required to account for the flexible pronunciation changes in multi-accent speech.

3. Multi-Accent Modeling

To model each accent individually, we use iterative dynamic programming (DP) alignment to "recognize" accent-specific units automatically from accented speech data. The phone models were bootstrapped from standard speech. We apply the likelihood ratio test to obtain "reliable" accent-specific units. This accent unit generation process is performed on each regional accent data. Finally, the original standard speech model is reconstructed into a multi-accent model through the use of accent-specific unit models to cover pronunciation changes.

3.1. Accent-specific units

We use automatic phone recognition with free grammar together with DP alignment to generate a set of accent-specific units. In phone recognition, the decoding formula is

$$B^* = \arg\max_B P(B)P(X \mid B) \quad (1)$$

where B is the canonical phone sequence, and X is the input speech vectors. Due to accent effects, some standard initial and final units cannot be pronounced correctly. Eq.1 needs to be rewritten by taking accented changes into consideration. Suppose a certain initial or final unit is pronounced in several ways in different accents. The output phonetic sequence including alternative representations forms surface form sequence S. Then the decoder formula becomes

$$B^* = \arg\max_B \left[P(B) \sum_S P(X \mid B,S) P(S \mid B) \right] \quad (2)$$

where $P(B)$ is the language model, $P(X \mid B,S)$ is the acoustic model, and $P(S \mid B)$ is the pronunciation model. In general, the acoustic model generation procedures assumes that

$$P(X \mid B,S) = P(X \mid B) \quad (3)$$

That means the acoustic model is based on canonical transcriptions and standard speech. If surface form transcriptions are available, the acoustic model training can be expressed as

$$P(X \mid B,S) = P(X \mid S) \quad (4)$$

which means the acoustic model is based on alternative transcriptions and accented speech. Obviously, both $P(X \mid B)$ and $P(X \mid S)$ are sub-optimal acoustic models if accent changes are considered. Ideally, both of them should be taken into account for acoustic model generation. In Eq.2, the combination of B and S is called an *accent-specific unit* and $P(X \mid B,S)$ is called the *accent-specific model*.

The DP alignment is performed iteratively on different sets of accented speech, to generate accent-specific units as shown in Fig. 1. We use likelihood ratio test as a confidence measure to select the more reliable accent-specific units [3],

Cantonese and Wu accented data are put into phone recognition individually to obtain the respective accent-specific unit sets. There are 18 selected accent-specific units for initials in Cantonese-accented speech, and 16 selected ones for Wu-accented speech. Note that although some pairs occur in both cases of Cantonese and Wu accented data (e.g., $zh \rightarrow z$), the tendency of the change and the corresponding acoustic parameters are distinct for different accents.

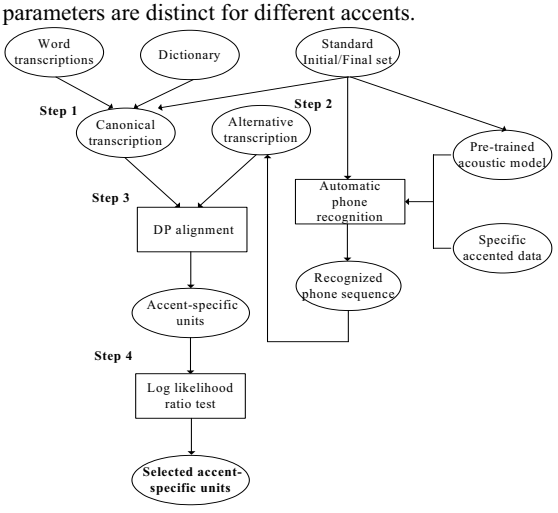

Fig. 1: The procedures of generating accent-specific units.

3.2. Auxiliary accent trees

We use decision tree based state tying for context-dependent triphone models [9] in our system. The selection of the question set, the size of decision trees and the selection of central units are key issues in decision tree based state tying.

In our system, the structure of triphones of accent-specific units is similar to that of standard triphones except for the central unit. The former is an accent-specific unit, and the latter is a single initial/final unit. Decision trees for accent-specific triphone units are called *auxiliary accent trees* as opposed to *standard decision trees* of standard triphones. Compared to standard decision trees, auxiliary trees are also phonetic binary trees in which a yes/no question is attached to each node. On the other hand, the question set for these accent trees is enlarged to include the accent-specific units. The tree size is smaller than that of standard decision trees due to fewer training samples.

The parameters of accent-specific triphone units are initialized from original pre-trained context-independent initial/final models b_i and context-independent accent-specific unit model $b_i\_s_i$, and are re-trained using different transcriptions.

$$\Phi'(b_i) \rightarrow \Phi(b_i\_s_i) \quad (5)$$

$$\Phi(b_i\_s_i) \rightarrow \Phi'(b_i\_s_i) \quad (6)$$

$$\Phi'(b_i\_s_i) \rightarrow \Phi(b_{i-1}, b_i\_s_i, b_{i+1}) \quad (7)$$

$$\Phi(b_{i-1}, b_i\_s_i, b_{i+1}) \rightarrow \Phi'(b_{i-1}, b_i\_s_i, b_{i+1}) \quad (8)$$

where Φ is the starting HMM model set and Φ' is the re-estimated HMM model set. Initial parameters of $\Phi(b_i\_s_i)$ are cloned from $\Phi'(b_i)$ and re-estimated using the BW algorithm with transcriptions in terms of standard initial final units as well as accent-specific units. Similarly, the initial model of $\Phi(b_{i-1}, b_i\_s_i, b_{i+1})$ is cloned from $\Phi'(b_i\_s_i)$ and re-estimated with triphone transcriptions.

3.3. Reconstruction of acoustic models

An individual set of auxiliary accent trees is constructed for each specific accent (e.g. Cantonese or Wu). The structure, the shape and the size of the auxiliary accent trees are unique to each accent even when the central unit is shared between different accent groups (e.g., when the central unit 'zh_z' appears in both accents).

Leaf nodes of decision tree represent tied-states in tree-based clustering ASR system. Hence, acoustic model reconstruction is equivalent to tree merge between auxiliary trees and standard decision trees as described in Fig.2. Determined by the minimum distance measure between tied states, leaf nodes from auxiliary accent trees are merged into the relevant modes of standard decision trees. More than one auxiliary accent trees can be attached to one standard decision tree, representing different accent changes. Following this tree merge the pre-trained acoustic model is reconstructed, and includes Gaussian mixture distributions from the standard model as well as those "borrowed" from accent-specific triphone models. As a result, the structure and the shape of the Gaussian distribution are adjusted. Additional Gaussians from tied states of the auxiliary trees are moved to the distribution boundaries to cover pronunciation variations from multiple accents.

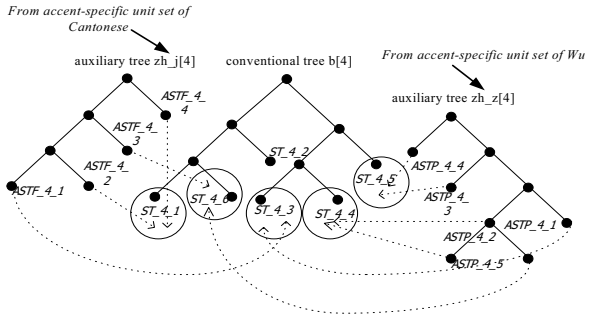

Fig. 2: Model reconstruction through decision tree merge.

In this approach, accent trees are only used during the state-tying procedure for accent-specific unit models, but not in the model estimation and decoding steps. That is, the structure of standard decision trees, the decoding complexity as well as decoding dictionary and recognizer are unchanged. There is no modification in the decoding procedure. More importantly, each auxiliary accent tree is uniquely attached to a single standard decision tree, so no model confusion is introduced. On the other hand, several different auxiliary trees can be attached to one standard decision tree to cover more types of accent changes. Moreover, since the model reconstruction is performed at the state level, a wide range of accent changes which cannot be represented at the phone level are thereby covered.

4. Recognition Experiments

We evaluate our approach in a Chinese telephony short phrase recognition task. There is no word n-gram in these short phrases so that we can isolate the effect of our approach without the influence from higher level information.

All speech data were sampled at 8 kHz and 8 bit-rate. The baseline acoustic model was trained using 50 hours of Putonghua speech including 100 speaker's utterances (each speaker has 200 utterances). The HMM topology is three-states, left-to-right without skips. The acoustic features are $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. Twenty-one standard initials and 38 finals were used to generate context-independent HMMs. We used HTK decision tree based state tying procedures to build 12 Gaussian-component triphone models with 5500 tied-states. 2000 continuous utterances with 23,685 syllables from 20 Cantonese-accented speakers (Dev1) and 20 Wu-accented speakers (Dev2) were used to extract two sets of accent-specific units. The test data consists of three parts: Test1 includes 900 Cantonese-accented utterances from 9 speakers (4 females and 5 males), excluded from Dev1; Test2 similarly consists of Wu-accented utterances; Test3 is used for performance comparison and consists of 900 Putonghua utterances selected from 9 native speakers. Speakers were instructed to speak the same phrases in these three sets.

Using log likelihood ratio test as confidence measure, 45 and 39 accent-specific units for Cantonese and Wu accented speech were extracted from the initial 8756 mapping pairs from Dev1 and Dev2 sets. Using the HTK state clustering approach, we constructed 135 auxiliary accent trees with 875 tied-states

and 117 auxiliary accent trees with 613 tied-states for Cantonese and Wu accent-specific triphone units, respectively. Through model reconstruction, all the auxiliary accent trees merged to the pre-trained standard model with 5500 tied states of 177 standard decision trees. The reconstructed model included 83,856 Gaussians and each state has 15.2 Gaussians on average. To make a fair comparison, we generated an enhanced acoustic model with 5500 tied-states and 15 Gaussian-component per state. The recognition performances are compared in Table 2.

System	Word Error Rate (WER) %		
	Test1 Can. accent	Test2 Wu accent	(Test3) Standard speech
Baseline	20%	12.7%	7.9%
Baseline HMM + Pronunciation dictionary	17.9% (-2.1)	11.1% (-1.6)	7.6% (-0.3)
Enhanced HMM with 15 Gau. per state	18.6% (-1.4)	12.4% (-0.3)	7.5% (-0.4)
Baseline HMM with MAP using Dev1	15.1% (-4.9)	15.8% (+3.1)	15.7% (+6.8)
Baseline HMM with MAP using Dev2	24.2% (+4.2)	9.1% (-3.6)	14.1% (+6.2)
Reconstructed HMMs	**15.6% (-4.4)**	**9.4% (-3.3)**	**7.2% (-0.7)**

Table 2: Lower WER for using reconstructed model compared to using MAP adaptation and augmented dictionary

Table 2 shows that accents have a great adverse impact on recognition accuracy if the acoustic model is trained from just the standard speech data. The second system shows that augmented dictionary with multiple pronunciations [2, 4] can cover some accented changes and gives an absolute 2.1% and 1.6% WER reduction on Test1 and Test2. More significantly, we can see that the reconstructed acoustic model gives a significant 4.4% and 3.3% absolute WER reduction compared to the baseline, and an additional 3% reduction with respect to using enhanced HMM at the same model complexity. This result indicates that the adjusted mixture distribution structure of the reconstructed model is robust enough to cover accent changes at different levels. On the other hand, directly increasing Gaussian components in the model results in poor estimation of some Gaussians when the training data is limited. In another experiment, we used Dev1 and Dev2 as adaptation data, and applied MAP adaptation on the baseline model. This approach gives a good 4.9% and 3.6% WER reduction on Cantonese and Wu accented speech, but led to serious performance degradations (6.8% and 6.2% WER increase) on native Putonghua speech. The adapted model using Dev1 Cantonese-accented data also leads to 3.1% performance degradation tested on Wu-accented speech, and vice versa. Through MAP adaptation, the acoustic parameters are irreversible changed to cater for accented speech and are no longer suitable for Putonghua speech. However, our reconstructed model includes its own Gaussians from standard speech acoustic model as well as those borrowed from different accent-specific unit models. The borrowed Gaussians are used only to adjust the structure of original mixture distribution and not to change the parameters. By jointly using these Gaussian distributions we are able to cover the diversity in multiple accents without sacrificing the performance on standard speech.

5. Conclusions

We have described an approach of acoustic model reconstruction for multi-accent speech recognition. In order to differentiate between multiple accents, different sets of accent-specific units were generated individually based on data-driven method. We generated auxiliary accent trees for accent-specific triphone units, and merge them with standard speech decision trees for acoustic model reconstruction under a context-dependent triphone framework. This approach adjusts the structure of the original mixture distribution in the standard speech models without changing model parameters, thereby improving model robustness and resolution to cover accent variability. Experimental results show that our proposed approach provides a significant 4.4% and 3.3% absolute WER reduction for Cantonese and Wu accented speech, a superior performance to other approaches. Compared to MAP adaptation, our method performs equally well on a particular accent without sacrificing the performance on other accents. Our approach is applicable to speaker-independent systems handling multiple and mixed accents.

6. References

[1] W.K. Liu and P. Fung, "Fast accent identification and accented speech recognition", In *Proc. ICASSP1999*, Mar. 1999, Phoenix, USA.

[2] Ch. Huang et.al., "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition", In *Proc. ICSLP2000*, Oct. 2000.

[3] Y. Liu and P. Fung, "Partial change accent models for accented Mandarin speech recognition", In *Proc. ASRU2003*, Dec. 2003.

[4] M. K. Liu, et.al., "Mandarin accent adaptation Based on context-inependent/context-dependent pronunciation modeling", In *Proc. ICASSP2000*

[5] L.M. Tomokiyo, "Recognzing non-native speech: characterizing and adapting to non-native usage in LVCSR", Ph.D thesis, Carnegie Mellon University, 2001

[6] Z. Wang, T. Schultz, A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech", In *Proc. ICASSP2003*, Apr. 2003

[7] Katagiri, S., et.al., "Pattern recognition using a family of designed algorithm based upon the generalized probabilistic descent method", Proc. IEEE 86, pp2345-2373, 1998.

[8] J. H, Huang, *Chinese Dialects*, XiaMen University Press, XiaMen, 1987 (Chinese version).

[9] Young, et.al., *The HTK book*, Entropic Cambridge Research Laboratory, 1999

[10] P. Fung, S. Huang, D Graff, *HKUST Mandarin Telephone Speech Part I and II*, http://www.ldc.upenn.edu/, 2005.