Effects and modeling of phonetic and acoustic confusions in accented speech

Pascale Fung[a)] and Yi Liu
Human Language Technology Center, Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

(Received 22 July 2004; revised 28 June 2005; accepted 21 July 2005)

Accented speech recognition is more challenging than standard speech recognition due to the effects of phonetic and acoustic confusions. Phonetic confusion in accented speech occurs when an expected phone is pronounced as a different one, which leads to erroneous recognition. Acoustic confusion occurs when the pronounced phone is found to lie acoustically between two baseform models and can be equally recognized as either one. We propose that it is necessary to analyze and model these confusions separately in order to improve accented speech recognition without degrading standard speech recognition. Since low phonetic confusion units in accented speech do not give rise to automatic speech recognition errors, we focus on analyzing and reducing phonetic and acoustic confusability under high phonetic confusion conditions. We propose using likelihood ratio test to measure phonetic confusion, and asymmetric acoustic distance to measure acoustic confusion. Only accent-specific phonetic units with low acoustic confusion are used in an augmented pronunciation dictionary, while phonetic units with high acoustic confusion are reconstructed using decision tree merging. Experimental results show that our approach is effective and superior to methods modeling phonetic confusion or acoustic confusion alone in accented speech, with a significant 5.7% absolute WER reduction, without degrading standard speech recognition. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.2035588]

PACS number(s): 43.72.Ne, 43.72.Ar, 43.71.An, 43.71.−k [DOS] Pages: 3279–3293

I. INTRODUCTION

Most state-of-the-art automatic speech recognition (ASR) systems fail to perform well when the speaker has a regional accent different from that of the standard language the systems were trained on. The high error rate is largely due to the effects of phonetic confusions and acoustic confusions in accented speech. Previous studies on accented speech recognition investigated in detail the effect of phonetic confusions (Liu *et al.*, 2000; Tomokiyo, 2001) or acoustic confusions (Huang *et al.*, 2000). However, the distinction and correlation between phonetic and acoustic confusions, in particular how to model their different roles for better ASR performance, are less clear. We suggest that it is essential to distinguish phonetic and acoustic confusions, as well as understand their relationship and roles in accented speech in order to achieve better recognition performance.

Phonetic confusions in accented speech are caused by the speaker pronouncing an expected phone in a different way (for example, when /zh/ is pronounced as /z/). A phone is the fundamental sound category that is represented by a particular group of articulatory features found in languages (Stevens, 1998). In the speech production process, a speaker first retrieves the canonical pronunciation of the word from his/her mental lexicon in terms of phoneme sequence (i.e., baseform), and then forms the articulatory shape of the pronunciation in terms of phones (i.e., surface form). Due to the effect of different accents and pronunciation habits, the surface form production can be different from that of the baseform. From this point of view, phonetic confusion can be regarded as the probabilistic transformation from a baseform unit to a surface form unit. In speech recognition, it is the erroneous recognition of a baseform phone into a different surface form phone.

On the other hand, acoustic confusion arises when the accented speech is found to lie acoustically somewhere between two baseform phones and can be equally recognized as either (for example, when it is in between /zh/ and /z/). Acoustic confusion can also come from data and recognizer-related confusions (Strik and Cucchiarini, 1999; Fung *et al.*, 2000), in addition to pronunciation variation.

In this paper, we focus on accent-specific phonetic and acoustic confusions. Phonetic and acoustic confusions are common and amorphous in accented speech, which degrade the recognition performance if they are not well accounted for.

A common approach to reduce phonetic confusion in ASR is by extending the phone set and generating a dictionary with multiple pronunciations (Bacchiani and Ostendorf, 1998; Chen *et al.*, 2002, Li *et al.*, 2000). In this approach, phonetic set is extended to include more surface form variants by either using hand-defined symbols based on phonological knowledge (Li *et al.*, 2000) or by using data-driven methods (Bacchiani and Ostendorf, 1998; Chen *et al.*, 2002). For example, Li *et al.* (2000) used pre-defined SAMPA-C symbols to differentiate Wu accented Mandarin pronunciations for spontaneous speech annotation. Chen *et al.* (2002) applied the chi-square test to design additional phonetic units for phonetic confusions with short duration. Bacchiani and Ostendorf (1998) proposed a data-driven method to generate

[a)]Electronic mail: pascale@ee.ust.hk

acoustic subword unit (ASU) to capture phonetic confusions. Jurafsky *et al.* (2001) showed that triphones are a good phone set for modeling multiple pronunciations.

Moreover, augmenting the pronunciation dictionary by pronunciation variations typically found in accented speech provides more hypotheses in the decoder search space which sometimes leads to better recognition results. Huang *et al.* (2000) and Liu *et al.* (2000) established accent-related pronunciation dictionaries, where the alternatives in the dictionaries are learned from accented speech data. Liu and Fung (1999) generated accent-adapted dictionary using some supra-segmental information to model the phonetic confusions in Cantonese-accented English.

To model acoustic confusions, especially those in accented speech, a commonly adopted method is to modify the acoustic parameters to cover accent variations. For example, retraining acoustic models using a large amount of accented speech data (Huang, *et al.*, 2000, Liu *et al.*, 2000), applying maximum *a posteriori* (MAP) or maximum log likelihood ratio (MLLR) on speaker-independent models to adapt to the acoustic characteristics of a particular accent (Young, 1999; Tomokiyo, 2001). Juang and Katagiri (1992), Chou *et al.* (1992) and Katagiri *et al.* (1998) proposed a discriminative training approach that uses the local error information to refine the acoustic model. Recently, Nakamura (2002) proposed restructuring Gaussian mixture density functions with Gaussian mixture sharing to restore local modeling mismatch within the confused acoustic models.

However, it is not sufficient to either model phonetic confusion or acoustic confusion exclusively as in the abovementioned methods, when phonetic and acoustic confusions are correlated but different.

Simply using extended phonetic units is insufficient in reducing a lot of phonetic confusions which also include acoustic confusions. Even though an accent-specific dictionary with multiple pronunciations provides a larger hypothesis space to cover phonetic confusions, a larger search space also leads to more lexical confusion *if* the underlying models are acoustically confusable. In other words, the increase in dictionary size can increase recognition errors if the underlying models already contain acoustic confusions. Meanwhile, the local error information used for discriminative training to reduce acoustic confusions is based on recognition errors, which may be attributed to various recognizer and data design configurations, not just because of accent. In addition, retraining and using MAP or MLLR adaptation of acoustic models lead to irreversible changes in acoustic parameters that are not suitable for native speech recognition. This results in performance degradation in speaker-independent systems.

In this paper, we propose methods to measure phonetic and acoustic confusions and reduce them for optimal speech recognition performance on accented speech without sacrificing the performance on standard speech. The paper is organized as follows. In Sec. II, we analyze a special case of Cantonese-accented Mandarin speech, which is used as our test case. Section III outlines the distinction and correlation between phonetic and acoustic confusions in accented speech. Section IV describes the mechanism of reducing both the phonetic and acoustic confusions using accent-specific units and acoustic model reconstruction. In Sec. V, experiments on accented Mandarin telephony speech are presented. We summarize our work and present our conclusions in Sec. VI.

II. CANTONESE ACCENT IN MANDARIN

Accent is a more serious problem for native Mandarin speakers than for native speakers of most other languages. In addition to the standard Chinese Mandarin (also known as Putonghua) spoken by radio and TV announcers, there are seven major language regions in China, including Guanhua, Wu, Yue, Xiang, Kejia, Min, and Gan (Huang, 1987). These major languages can be further divided into more than 30 sub-categories of dialects. In addition to lexical, syntactic, and colloquial differences, the phonetic pronunciations of the same Chinese characters are quite different between Mandarin and the other Chinese languages. Only 70% of Chinese speakers on Mainland China are native speakers of Guanhua, the language group most related to Mandarin. Among these, only a minority speak with the standard Mandarin accent. Consequently, accent distribution among Mandarin speakers can be as varied as that among European speakers of English. Cantonese is an important regional language and is spoken by tens of millions of speakers in south China, Hong Kong, and overseas. 60% of the pronunciations between Cantonese and Mandarin are not even close to each other (Huang, 1987). In this section, we focus on the phonetic and acoustic analysis of Cantonese and Mandarin, especially on the pronunciation differences of their subword units, to highlight the phonological differences between the two languages. Cantonese-accented Mandarin is used as the test case for our work in this paper.

In Chinese ASR systems, initial and final units are conventionally used as subword units instead of phonemic units. One initial corresponds to one phoneme, while one final may consist of one or several phonemes. Without taking into account tonal differences, there are 21 initials and 37 finals for Mandarin, compared to 19 initials and 53 finals in Cantonese (Lee *et al.*, 2002). Initials in both Mandarin and Cantonese consist of a single consonant. However, the initial inventories for these two languages are different. In contrast to Mandarin initials, Cantonese initials do not have retroflexed affricatives (e.g., /zh/, /ch/, /sh/, and /r/), but include one additional velar nasal /ng/. Table I gives an example of a comparison between Mandarin and Cantonese initials with respect to the place and method of articulation. The structure of Cantonese finals is more complicated than that of Mandarin. Cantonese finals have six different consonant codas (/m/, /n/, /ng/, /k/, /p/, and /t/) in contrast to the two codas /n/ and /ng/ in Mandarin finals. Cantonese finals have five categories: vowel, diphthong, vowel with nasal coda, vowel with stop coda and syllabic coda. On the other hand, Mandarin finals were comprised of a vowel or diphthong nucleus preceded by an optimal medial and followed by an optimal nasal.

Consequently, native Cantonese speakers often have difficulty pronouncing many basic Mandarin initials and finals.

TABLE I. Mandarin initials vs Cantonese initials.

Manner of articulation	Place of articulation	Mandarin initials
4 Plosive	Labial	b
	Alveolar	d
	Velar	g
Aspirated plosive	Labial	p
	Alveolar	t
	Velar	k
Affricates	Alveolar	z
	Retroflex	zh
	Dorsal	j
Aspirated affricates	Alveolar	c
	Retroflex	ch
	Dorsal	q
Nasals	Labial	m
	Alveolar	n
Fricatives	Labiodental	f
	Alveolar	s
	Retroflex	sh
		r
	Dorsal	x
	Velar	h
Laterals	Alveolar	l
Plosive	Labial	b
	Alveolar	d
	Velar	g
Aspirated plosive	Labial	p
	Alveolar	t
	Velar	k
Plosive, lip-rounded	Velar, labial	gw
Aspirated plosive, lip- rounded	Velar, labial	kw
Nasals	Labial	m
	Alveolar	n
	Velar	ng
Liquid	Lateral	l
Affricate, unaspirated	Alveolar	z
Affricate, aspirated	Alveolar	c
Fricative	Alveolar	s
	Dental-labial	f
	Vocal	h
Glide	Alveolar	j
	Labial	w

They use some of the typical strategies of language learners to compensate for such difficulties, including phonological transfer, overgeneralization, prefabrication, epenthesis, etc. For example, the pronunciation of the retroflexed affricative /zh/ is sometimes similar to that of the dental velar /z/ among Cantonese speakers. Since there is no /zh/ in the Cantonese initial set, the speaker naturally moves this pronunciation to the most similar initial unit /z/ from the Cantonese initial set. On the other hand, such pronunciation is distinct from the canonical pronunciation of /z/ since the speaker needs to distinguish the pronunciation between /zh/ and /z/. Sometimes, this intention to distinguish leads the Cantonese speaker to pronounce /zh/ as /j/. Since the speaker is trying to say /zh/ and not /z/ or /j/, the phonological transfers of “$zh \rightarrow z$” and “$zh \rightarrow j$” lead to confusable pronunciations which are correlated with yet different from the canonical pronunciations of /z/ and /j/. Moreover, this type of change is unidirectional in accented speech, i.e., there are no “$z \rightarrow zh$” or “$j \rightarrow zh$” transfers. The degree and tendency of confusions between “$zh \rightarrow z$” and “$z \rightarrow zh$”, and between “$zh \rightarrow j$” and “$j \rightarrow zh$” are quite different.

III. PHONETIC CONFUSION VERSUS ACOUSTIC CONFUSION

A. Phonetic confusions and acoustic confusions are different yet correlated

There are different types of phonetic and acoustic confusions in speech recognition systems. While some phonetic and acoustic confusion are correlated with each other, others are not, and whereas some are caused by accented speech, others are due to inadequacies and idiosyncrasies in the design and implementation of the recognizer or the training data. In this paper, we focus on analyzing and reducing phonetic and acoustic confusion caused by accented speech.

Phonetic confusion is a property of relating phone instances to acoustic models whereas acoustic confusion is a property of acoustic models. In accented speech, phonetic confusion is caused by the pronunciation of an expected phone into a different one whereas acoustic confusion arises from a pronounced phone lying between two standard phones acoustically (Liu and Fung, 2003a; Tsai and Lee, 2003). For a speech recognizer trained on standard speech, phonetic confusion is then the erroneous recognition of a phonetic unit in the accented speech into another phonetic unit in the standard speech. It can be regarded as the probability of the transformation from a baseform unit to a surface form unit. Acoustic confusion, on the other hand, is at a more fundamental level and describes the distance between the phonetic unit in accented speech and phonetic units represented by two baseform models, in terms of acoustic properties.

Phonetic and acoustic confusions are different yet correlated in the speech recognition task. If the acoustic models of two phonetic units are close to each other (i.e., not easily separable), then these models have low discriminative ability and will cause phonetic confusions in the final recognition task, irregardless of whether the input speech is accented or not. However, even if the trained acoustic models have good separability, accented speech might produce a phone that lies somewhere between two models and again cause acoustic confusion, resulting in phonetic confusion. In other cases, the accented speech might produce one phone that is clearly close to another, different phone in the standard speech. This causes phonetic confusion, even though there is no acoustic confusion between models.

We use Fig. 1 to illustrate the distinction and correlation between the phonetic confusion and the acoustic confusion. Suppose “A” is a phonetic unit and “B” is another phonetic unit that is often confused with “A.” Acoustic models for “A” and “B” consist of a single Gaussian component, $G_{\mathrm{A}}(\mu_{\mathrm{A}}, \sigma_{\mathrm{A}})$ and $G_{\mathrm{B}}(\mu_{\mathrm{B}}, \sigma_{\mathrm{B}})$, respectively, where μ and σ are the mean and the variance. The phonetic confusion between units “A” and “B” are measured using $P(\mathrm{B}|\mathrm{A})$ which is computed using occurrence frequencies (Byrne *et al.*, 2001;

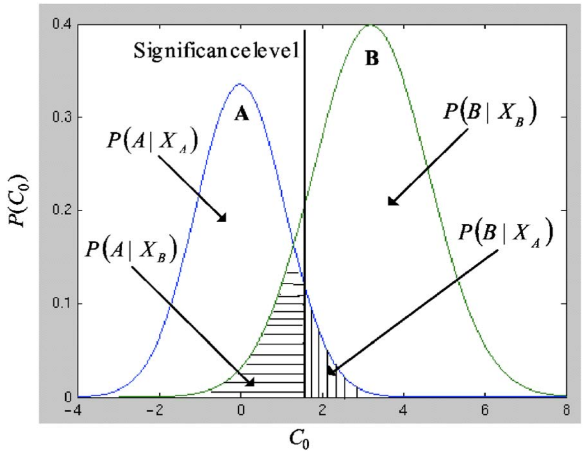

FIG. 1. An example of phonetic and acoustic confusions. C_0 is one dimension of the mean and $P(C_0)$ is the relevant output distribution.

Liu and Fung, 2003a; Chen *et al.*, 2002). The more "A" maps to "B," the higher the phonetic confusion. On the other hand, the acoustic confusion is measured using the acoustic distance between models "A"; and "B", i.e., the distance between the Gaussian components. This distance is computed using Gaussian distance measure (Li and King, 1999; Liu and Huang, 2000). In this case, the more model "A" overlaps with model "B" (the shaded area in Fig. 1), the higher the acoustic confusion between "A" and "B." Obviously, phonetic and acoustic confusions are measured differently.

Suppose the acoustic samples for phone A and phone B are X_{A} and X_{B}, respectively, in the accented speech. If the acoustic sample is located in the shaded area, it can be assigned to either $P(\mathrm{B}|X_{\mathrm{A}})$ or $P(\mathrm{A}|X_{\mathrm{B}})$, causing phonetic confusions. Conversely, acoustic confusion is caused by a large overlap between "A" and "B," causing the misclassification of $P(X_{\mathrm{A}}|\mathrm{B})$ and $P(X_{\mathrm{B}}|\mathrm{A})$.

Even if model A and model B do not have any overlap, X_{A} can still be recognized as phone B, if the accented speech differs from standard speech. In this case, there is phonetic confusion without acoustic confusion.

We wish to point out that there are other conditions that lead to phonetic confusions, even if the input speech has no accent: (1) models "A" and "B" are confusable even in standard speech. This corresponds to the underlying acoustic confusions (e.g., /b/ and /d/, /in/ and /ing/ in Mandarin speech). Representing this condition in Fig. 1 is the large shaded area, and the high overlap ratio. Hence, even if the speaker accurately pronounces "A" or "B," chances for mismatched outputs still exist; (2) models are poorly trained because of biased data or incorrect phonetic transcriptions due to transcriber disagreement.

Assuming the above two factors are held constant, i.e., we use the same set of training data and transcriptions and the same training methods for an ASR system, we are interested in studying how best to reduce phonetic and acoustic confusions due to the accent effect.

B. Measuring phonetic and acoustic confusions

1. Measuring phonetic confusions

Phonetic confusions are measured in terms of the *distribution* of the mapping between surface form and baseform phones. Due to the effect of accented speech, the baseform (standard speech) and surface form (accented speech) sequences of a word differ. For example, the word (China) has the standard pronunciation represented by the baseform sequence "zh ong g uo." Cantonese-accented Mandarin speech might produce different surface form representations such as "z ong g uo", "ch ong g uo" or "j ong g uo."

Aligning the baseform and surface form representations and counting the mapped phone pairs is an obvious way to estimate phonetic confusion distribution. However, as we mentioned in Sec. III A, this type of confusion can be caused by accent as well as the recognizer or training data design and implementation. As we need to focus on phonetic confusion caused by accent effect, it is necessary to impose a confidence measure on the phonetic confusion pairs. Intuitively, if a particular phone A in input speech is often misrecognized as phones B, C, D, etc., then we reason that the phone model A in the ASR system is unreliable either due to training data bias or recognizer design. Similarly, if we find multiple phones being misrecognized as B, then we have reason to believe that the phone model B is unreliable. However, if A and only A is consistently misrecognized as B, then we suspect that there is a phonetic shift from B to A in the accented speech. Of course, there might be additional acoustic confusion between A and B as well, which can then be measured using another measure described in the next section.

We use *likelihood ratio test* as a confidence measure to evaluate the phonetic confusions. We use dynamic programming to align the phone sequences in the accented speech with standard baseform phone transcriptions. For a baseform phone b which is misrecognized as s, we count the occurrences of b, s, and b_s in the aligned data, which are c_1, c_2, and c_{12}, respectively. We have the likelihoods:

$$p=\frac{c_2}{N},\quad p_1=\frac{c_{12}}{c_1},\quad p_2=\frac{c_2-c_{12}}{N-c_1}, \tag{1}$$

where N is the total number of the phonetic units in the training set. The log of the likelihood ratio λ is then defined as follows:

$$\begin{aligned}\log\lambda &= \log L(c_{12},c_1,p)+\log L(c_2-c_{12},N-c_1,p)\\ &\quad -\log L(c_{12},c_1,p_1)+\log L(c_2-c_{12},N-c_1,p_2),\end{aligned} \tag{2}$$

where $L(k,n,x)=x^k(1-x)^{n-k}$ is a binomial distribution. In general, we use $-2\log\lambda$ instead of λ in practice (Manning and Schütze, 1999). The phonetic confusion distribution is then described as

$$D_{\mathrm{ph}}(b,s)=\frac{C}{-2\log\lambda} \tag{3}$$

where C is a constant estimated from data. Equation (1) shows that the likelihood of phonetic confusion depends not only on the occurrence frequency of b_s but also on the occurrence frequencies of b and s. Thus, we can distinguish between whether the models b and s are simply badly trained or there is indeed a phonetic shift from b to s. Moreover, since c_{12} differs from c_{21} (e.g., the occurrence number of

/zh/->/z/ is different from that of /z/->/zh/), this phonetic confusion distribution is asymmetric and unidirectional, in accordance with phonological knowledge about accented speech.

2. Measuring acoustic confusions

The degree of acoustic confusions can be measured by the dissimilarity or distance between two speech vectors, between a speech vector and a speech model, and between two speech models. For accented speech, we are interested in measuring the statistical dissimilarity between that of the accented speech model and the standard speech model. Common distance measures include Euclidian distance, Mahalanobis distance, Kullback-Leibler distance, etc. (Hwang, 1993; Liu and Huang, 2000). These measures assume that the distance between two vectors or models is symmetric, i.e., the acoustic distance from model "A" to model "B" is equal to that of from model "B" to model "A." However, it is well known that acoustic confusions in accented speech are asymmetric and unidirectional[1] (Liu and Fung, 2003a, Tsai and Lee, 2003). For speech recognition tasks, we need an asymmetric distance measure between continuous hidden Markov models (CHMM) with variable, multiple Gaussian components.

Tsai and Lee (2003) proposed an asymmetric acoustic distance measure that uses an asymmetric form of Mahalanobis distance, which is the averaged distance over all M mixtures and over all N states of two HMMs:

$$D_{ac}(\lambda_i,\lambda_j)=\sum_{s=1}^{N}\sum_{m_{i,s}=1}^{M} w_{m_{i,s}}\sum_{m_{j,s}=1}^{M} w_{m_{j,s}}d(g_{m_{i,s}},g_{m_{j,s}}).$$

However, the above distance measure simplifies multiple mixtures into one mixture before obtaining the *average* distance. This averaged distance sometimes does not correspond to true model distance as had been pointed out in previous research (Liu and Huang, 2000). Instead, we start from the method of parametric distance metric for mixture probability distribution function (PDF) described in (Liu and Huang, 2000), and propose an *asymmetric acoustic distance measure* for CHMM with multiple states and variable, multiple Gaussian components using a weight matrix. Suppose λ_i and λ_j are two different CHMM phonetic models which consist of N states. Each individual state is represented by a PDF representing multiple Gaussian components. Consider two different states s_{in} and s_{jn} of model λ_i and model λ_j,

$$s_{in}=\sum_{k=1}^{K} w_{in,k}g_{in,k}(\mu_k,\sigma_k)$$

and

$$s_{jn}=\sum_{l=1}^{L} w_{jn,l}g_{jn,l}(\mu_l,\sigma_l), \tag{4}$$

where $w_{in,k}$ and $w_{jn,l}$ correspond to the mixture weights of the kth and lth Gaussian components which satisfy $\Sigma_{k=1}^{K}w_{in,k}=1$ and $\Sigma_{l=1}^{L}w_{jn,l}=1$. According to the parametric distance metric for PDF, the distance $D(s_{in},s_{jn})$ is defined as

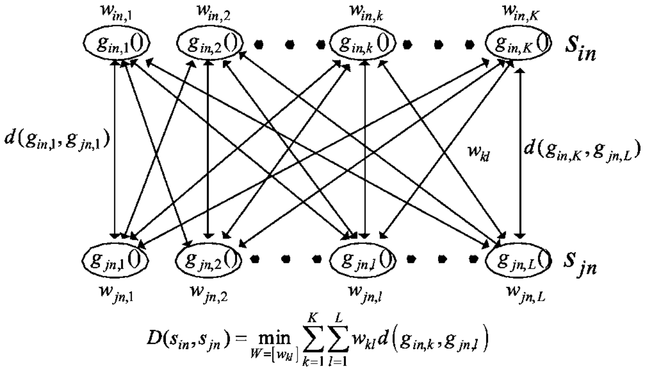

FIG. 2. Asymmetric distance measure for CHMM with multiple states and multiple Gaussian components.

$$D(s_{in},s_{jn})=\min_{W=[w_{kl}]}\sum_{k=1}^{K}\sum_{l=1}^{L} w_{kl}d(g_{in,k},g_{jn,l}), \tag{5}$$

where $W=[w_{kl}]$ is a weight matrix to be estimated by using a linear programming procedure, such as the Simlex tableau method (Cover and Thomas, 1991). $d(g_{in,k},g_{jn,l})$ is an element distance between two single Gaussian components. In order to consider the asymmetric property of acoustic confusions in accented speech, we use the asymmetric form of Mahalanobis distance:

$$d(g_{in,k},g_{jn,l})=(\mu_k-\mu_l)^T\sigma_2^{-1}(\mu_k-\mu_l).$$

The weight matrix $W=[w_{kl}]$ is determined under the constraints shown in Eq. (6):

$$\begin{aligned}
&w_{kl}\geqslant 0\\
&\sum_{k=1}^{K} w_{kl}=w_{jn,l}\quad 1\leqslant k\leqslant K,\\
&\sum_{l=1}^{L} w_{kl}=w_{in,k}\quad 1\leqslant l\leqslant L,\\
&\sum_{k=1}^{K} w_{in,k}=1,\\
&\sum_{l=1}^{L} w_{jn,l}=1.
\end{aligned} \tag{6}$$

The overall distance $D(s_{in},s_{jn})$ is then determined according to the weight matrix $W=[w_{kl}]$ and element distance between each Gaussian component, as described in Fig. 2.

Finally, the overall distance between model λ_i and model λ_j is calculated as a sum of each individual state distances:

$$D(\lambda_i,\lambda_j)=\sum_{n=1}^{N} D(s_{in},s_{jn}). \tag{7}$$

In this paper, we assume state alignment between two HMMs since the baseform and surface form models in accented speech have the same number of states. Equation (7) could be replaced by frame-to-state alignment if the models have different state numbers (Liu and Fung, 2001). Our proposed asymmetric acoustic distance measure is both computationally efficient and linguistically motivated. Owing to the different mixture weight matrices $W=[w_{kl}]$ and $W=[w_{lk}]$, as well as the asymmetric element distance, the distance $D(s_{in},s_{jn})$ is distinguished from $D(s_{jn},s_{in})$. Hence, our acoustic distance measure captures the fact that the acoustic confusion is asymmetric and unidirectional in accented speech.

Given the above-noted quantitative measures of acoustic and phonetic confusions, we can describe different classes of confusion in accented speech.

C. Combinations of phonetic and acoustic confusions in accented speech

There are four combinations of acoustic and phonetic confusions in speech recognition systems: (1) phonetic confusions and acoustic confusions are both low; (2) phonetic confusion is low and acoustic confusion is high; (3) phonetic confusion is high and acoustic confusion is low; and (4) phonetic confusions and acoustic confusions are both high.

Ideally, the subword units (e.g., phonemes and phones or initials/finals in Mandarin speech) used in ASR systems should be modeled and trained so that phonetic and acoustic confusions are both low for good discriminative-ness. Condition (1) is therefore desirable for ASR systems.

Condition (2) in which phonetic confusion is low but acoustic confusion is high is relatively rare. It happens when two phoneme models are acoustically confusable (i.e., with overlapping acoustic characteristics such as between /l/ and /n/), but accented speaker tends to distinguish the two phones very clearly, even more so than standard speakers (for example, Cantonese speakers never pronounce /l/ close to /n/). This type of confusion exists when native models are acoustically confusable (e.g., “l” and “n” in Mandarin) whereas accented speakers, by overcompensation, can separate the two pronunciations better than native speakers in their pronunciation (e.g., Cantonese speakers of Mandarin) (Huang, 1987). Under condition (2), accented speech does not adversely affect speech recognition performance. Example phone pairs in condition (2) for Cantonese-accented Mandarin are shown in the following:

1. $n \rightarrow l$, 2. $d \rightarrow p$, 3. $h \rightarrow k$, 4. $ei \rightarrow en$,
5. $ei \rightarrow ui$, 6. $ang \rightarrow iang$, 7. $c \rightarrow z$, 8. $k \rightarrow g$.

Condition (3) under which phonetic confusion is high and acoustic confusion is low indicates that phonetic confusion in this case is not caused by acoustic confusion, since acoustic models under this condition have good discriminative abilities. Accent is a predominant factor leading to phonetic confusion in this case. For instance, acoustic confusion between models /f/ and /x/ is low since there is little overlapping acoustic characteristic between standard Mandarin models of these two sounds. On the other hand, there is high phonetic confusion between /f/ and /x/ in Cantonese-accented Mandarin speech. In Cantonese-accented Mandarin speech, we have detected the following phone pairs that have high phonetic confusion but low acoustic confusions:

1. $ai \rightarrow uai$, 2. $h \rightarrow u$, 3. $ao \rightarrow ou$, 4. $t \rightarrow sh$,
5. $en \rightarrow iang$, 6. $sh \rightarrow r$, 7. $d \rightarrow zh$, 8. $x \rightarrow t$,
9. $j \rightarrow d$, 10. $h \rightarrow q$, 11. $d \rightarrow z$, 12. $ia \rightarrow e$,
13. $d \rightarrow n$, 14. $x \rightarrow t$, 15. $f \rightarrow sh$, 16. $n \rightarrow sil$,
17. $d \rightarrow m$, 18. $j \rightarrow b$, 19. $f \rightarrow z$, 20. $e \rightarrow uo$,
21. $l \rightarrow d$, 22. $x \rightarrow i$.

Under condition (4), phonetic and acoustic confusions are both high. If most of the phonetic units are phonetically and acoustically confusable, then perhaps the unit inventory is not well defined and/or the acoustic models are not well trained. The acoustic models do not have good separability and ASR performance will suffer greatly. Another factor is again accent. In most cases, the two factors co-exist. That is, the acoustic models do not have good separability *and* the accented speech differs from standard speech.

More important, accent effect is a key contributing factor to high acoustic and phonetic confusions. For example, the articulatory features of the retroflexed affricative /zh/ are similar to those of the dental velar /z/ for Cantonese-accented speakers. Since there is no /zh/ sound in the native Cantonese initial set, Cantonese speaker naturally shifts this pronunciation to the most similar initial unit /z/, found in native Cantonese phone set. However, the pronounced /zh/ by Cantonese speaker is not exactly /z/ either, but acoustically somewhere in between /zh/ and /z/. This shift leads to phonetic confusion as well as acoustic confusion between /zh/ and /z/ in Cantonese-accented Mandarin speech. An analysis of Cantonese-accented Mandarin speech data shows us that this type of confusion is limited to a particular set of subword units, such as the retroflexed affricatives to dental velars in Cantonese-accented Mandarin speech.

To illustrate the above, we plot the two-dimensional projection of the acoustic distribution of actual MFCC samples of accented versus standard Mandarin for the baseform /zh/ in Fig. 3. We can see that while “zh->z” and “zh->zh” share similar acoustic properties, “zh->j” and “zh-> others” are clearly different in terms of acoustic cluster shape and centroid.

Note that while our visualization method cannot show all of the parameters or variations of the original acoustic (LDA compression of the features may cause the loss of some variation information), it has been found that if two phenomena are dissimilar in two dimensions, they can only be more dissimilar in the original feature space (Peters and Stubley, 1998). In other words, the characteristics of partial changes in two dimensions are in accordance with their characteristics in higher dimensions. Examples of phone pairs which belong to condition (4) in Cantonese-accented Mandarin speech are shown below:

1. $ai \rightarrow an$, 2. $c \rightarrow ch$, 3. $c \rightarrow ch$, 4. $ch \rightarrow s$,
5. $d \rightarrow j$, 6. $f \rightarrow h$, 7. $h \rightarrow g$, 8. $in \rightarrow ing$,
9. $j \rightarrow x$, 10. $j \rightarrow zh$, 11. $m \rightarrow l$, 12. $q \rightarrow x$,
13. $s \rightarrow zh$, 14. $x \rightarrow z$, 15. $zh \rightarrow q$, 16. $zh \rightarrow s$.

Since accented speech only impacts ASR systems adversely in conditions (3) and (4), our task is to analyze and

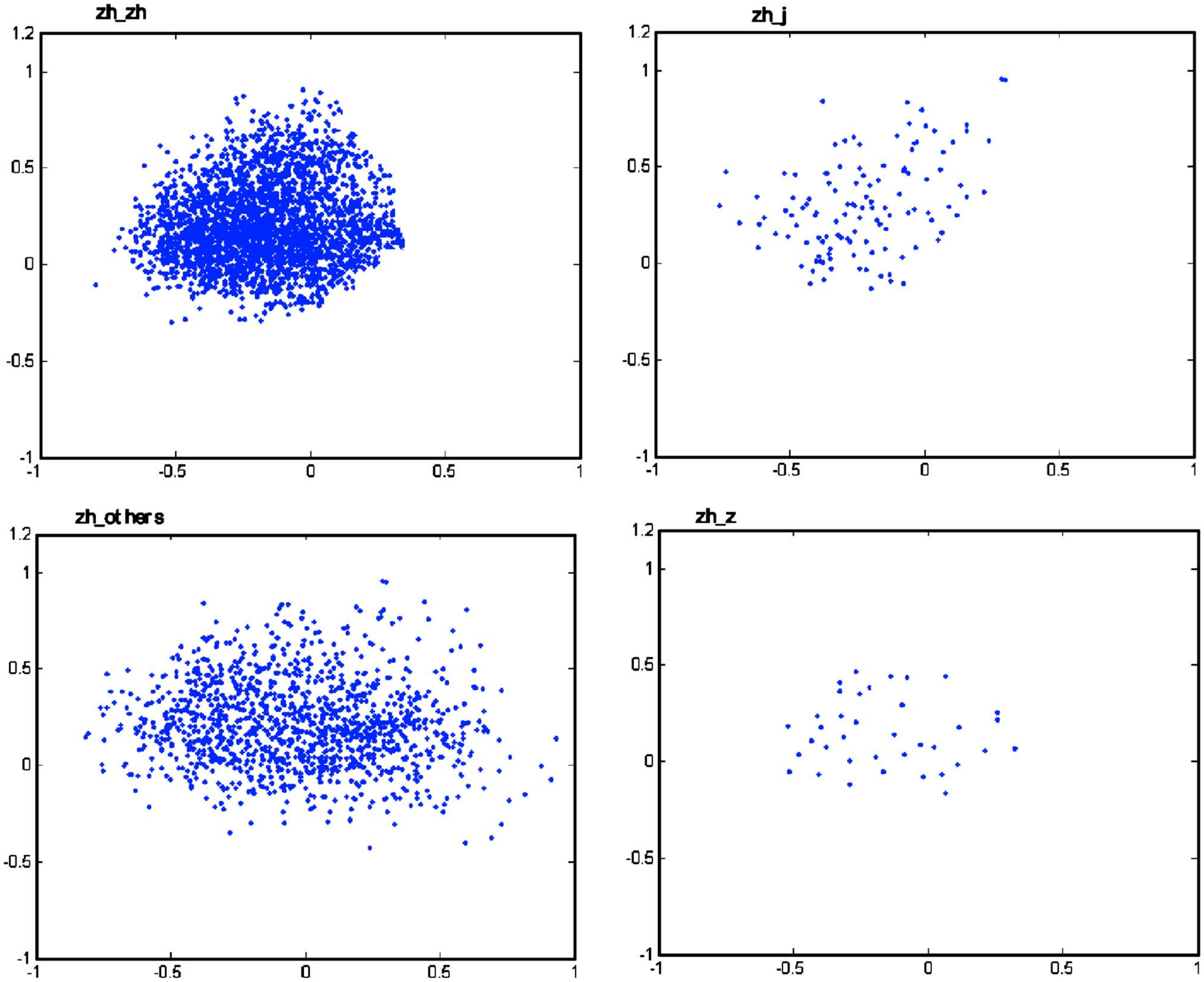

FIG. 3. Two dimensional MFCC samples of accented vs standard Mandarin for the baseform /zh/.

model accented speech with the objective of reducing phonetic and acoustic confusions under these conditions.

IV. REDUCING PHONETIC AND ACOUSTIC CONFUSIONS FOR ACCENTED SPEECH RECOGNITION

We studied four combinations of acoustic confusions and phonetic confusions in speech recognition. The investigation of these four combinations and the corresponding pronunciation phenomena in accented speech shows that the phonetic and acoustic confusions should be considered distinctively to improve recognition performance in accented speech recognition task. Figure. 4 gives examples of acoustic and phonetic distances of Chinese initials in the accent-specific units.

To model phonetic and acoustic confusions in accented speech for the task of speech recognition, we propose the following algorithm:

Modeling Phonetic and Acoustic Confusions in Accented Speech:

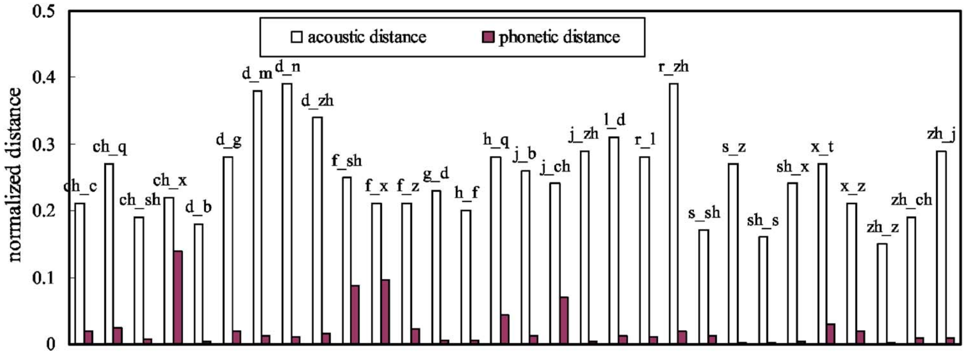

FIG. 4. Examples of normalized acoustic and phonetic distances in accent-specific units.

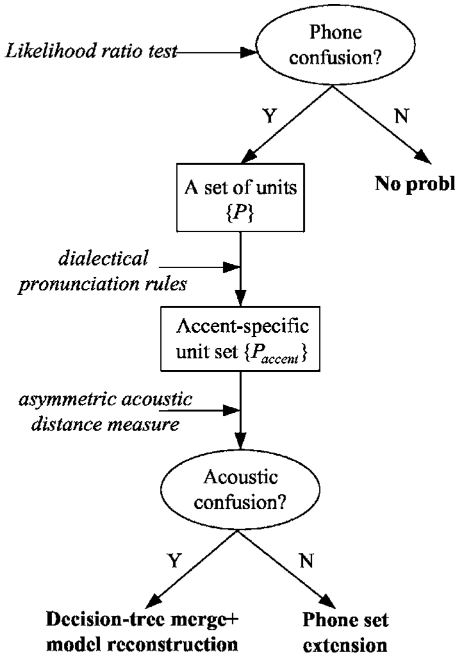
Likelihood ratio test
Phone confusion?
Y
N
A set of units {P}
No problem
dialectical pronunciation rules
Accent-specific unit set {P_accent}
asymmetric acoustic distance measure
Acoustic confusion?
Y
N
Decision-tree merge+ model reconstruction
Phone set extension

FIG. 5. State-transition charts of modeling phonetic and acoustic confusions in accented speech.

1. Identify phonetic confusion in the input speech by *likelihood ratio test* to generate a set of units $\{P\}$;
   1a. Identify accent-specific confusion pairs from $\{P\}$ by using dialectical pronunciation rules, and replace set $\{P\}$ by this new set $\{P_{\text{accent}}\}$.
2. Identify acoustic confusion from $\{P_{\text{accent}}\}$ using *asymmetric acoustic distance measure*, and form a set of units that have high phonetic confusion but low acoustic confusion $\{A\_l\}$ and another set of units with high phonetic confusion as well as high acoustic confusion $\{A\_h\}$;
3. For phonetic units in $\{A\_l\}$, form a multiple pronunciation dictionary with *extended phone set*;
4. For phonetic units in $\{A\_h\}$, use *acoustic model reconstruction* with decision-tree merging.

State-transition charts of the above-mentioned algorithms are shown in Fig. 5.

In the following sections, we first explain how to classify phone units into those with high and low phonetic and acoustic confusions in Sec. IV A. The algorithm for extending phone sets to form a multiple pronunciation dictionary units with high phonetic confusion but low acoustic confusion in $\{A\_l\}$ is described in Sec. IV B. The algorithm for acoustic model reconstruction with decision-tree merging for units with high phonetic and high acoustic confusion in $\{A\_h\}$ is detailed in Sec. IV C.

A. Classifying phone units according to accent effects

As we explained in Sec. III, only phonetic units with high phonetic confusion can lead to recognition errors, whether these confusions are caused by accented speech or other factors. Therefore the first step in modeling accented speech is to find phonetic units with high phonetic confusion during recognition. An initial set of baseform to surface form phone confusion pairs are found by dynamic programming alignment between the baseform and surface form transcriptions. The baseform transcription is a phoneme sequence corresponding to canonical pronunciations found in a standard pronunciation dictionary. The surface form transcription is a phone sequence with alternative pronunciation information, which can be obtained either by hand-labeled transcription or by a weighted finite-state transducer using a Classification and Regression Tree. In this step, we implement a flexible alignment tool that incorporates intersymbol comparison costs. These costs are based on phonetic feature distance between each pair of phone symbols, derived from linguistic rules (Fung *et al.*, 2000; Byrne *et al.*, 2001; Sproat, 2001).

Next, *likelihood ratio test* is applied to the DP-aligned baseform-surface form phone pairs to form a set of phonetically confusable units. As a result, 353 units are selected from the original 6573 initially found units. To help further distinguish between phonetically confusable units that are caused by accented speech from those caused by recognizer or data related factors, we use some linguistic rules to select a subset of the 353 units that are believed to be due to accented speech. For Cantonese-accented Mandarin, we apply the following linguistic rules in Cantonese dialectical pronunciations described in (Huang 1987):

(1) High confusions within retroflexed affricatives (e.g., /zh/, /ch/, /sh/ and /r/).
(2) High confusion between /f/ and /x/.
(3) One special velar nasal /ng/.
(4) Cantonese finals include /m/ coda.
(5) Pronunciation change in accented speech is unidirectional (e.g., /zh/ moves to /z/ and /r/ moves to /l/ but not vice versa).
(6) No medial in Cantonese finals.

These rules enable us to select 79 accent-specific units from the previous 353 pre-selected units for phonetic confusions.

These 79 phonetically confusable units are further divided into two classes: those with high acoustic confusions and the others with low acoustic confusions. The *asymmetric acoustic distance measure* is used to divide the units into high and low acoustic confusion pairs. 57 phone units are found to have high phonetic and high acoustic confusions whereas 27 phone units are found to have high phonetic but low acoustic confusions.

Having classified accent-specific phonetic units according to high and low acoustic confusions, we suggest selecting only phonetic units with low acoustic confusions to form alternate pronunciations and add into a pronunciation dictionary. For phone units with high acoustic confusions, we suggest that incorporating them into a pronunciation dictionary will further increase lexical confusions. Instead, we propose using decision tree merging with acoustic model reconstruction for this class of phone units.

B. Modeling phone units with high phonetic confusion and low acoustic confusion

When standard phonetic unit models are applied to accented speech recognition tasks, severe performance degra-

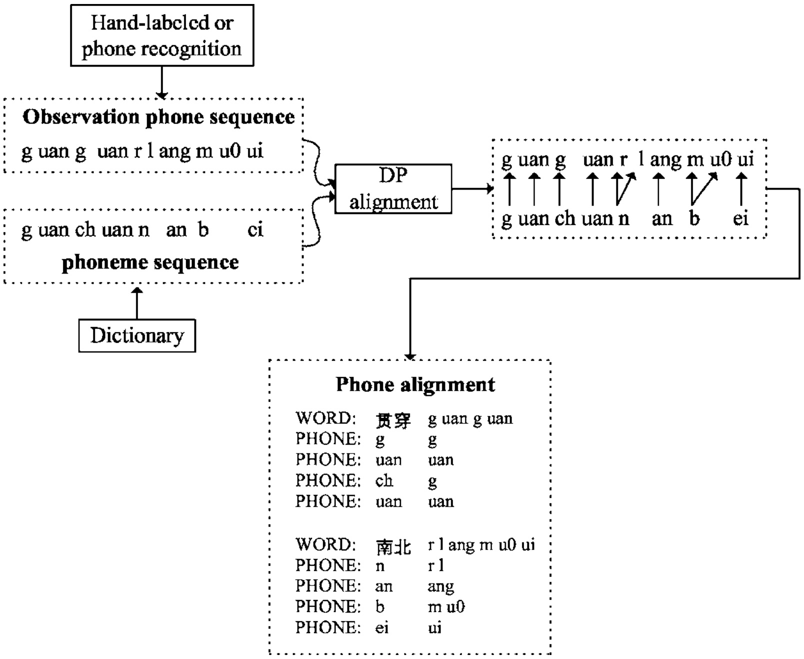

FIG. 6. Aligning baseform sequence to surface form sequence.

dation is observed (Huang, *et al.*, 2000). The increased phonetic variability in accented speech means the acoustic models of the defined units are not adequate for modeling such variability within subword units. Hence, we need to extend the original phonetic unit inventory to represent accented speech. The extended units are used to form alternate pronunciations in a pronunciation dictionary to cover phonetic variations (Holter, 1997 Riley *et al.*, 1999). Special attention must be paid to selecting units with low acoustic confusions. Our resultant multiple-pronunciation dictionary should cover only units with phonetic confusions but not those with acoustic confusions.

Adding these pronunciations, the dictionary is augmented and includes both standard initial/final units and accent-specific units. Compared to conventional multiple pronunciation dictionaries (Liu *et al.*, 2000; Huang *et al.*, 2000), our augmented dictionary uses selected units with low or no acoustic confusion. In other words, the use of such dictionary provides more chances for speech recognizer to output correct sequences without increasing lexical confusion. Moreover, pronunciation probabilities can be attached to each entry of the dictionary. These associated probabilities can be determined from training data using decision tree based structure as follows: A decision tree is constructed to predict the surface form of each reference phoneme by asking questions about its phonemic context. Each phoneme unit has a separate decision tree in which a yes/no question is attached to each node. These questions include information about the phoneme stream itself (such as stress, position, and the classes of neighboring phones), or the past output of the tree (including the identities of surface phones to the left of the current phone). From this, a probability distribution over the set of surface phone(s) for any given context can be determined by the alignment. The decision tree-based pronunciation model thus assigns probabilities to alternative surface form realizations of each phone depending on its context. When decision tree-based pronunciation modeling is carried out, it can be used to generate phone level networks to predict alternative pronunciations in terms of phone sequences. An example alignment is shown in the following Fig. 6.

C. Modeling phone units with high phonetic and high acoustic confusions

Due to high acoustic confusions and the resultant lexical confusions, the direct use of extended phone units to form alternative pronunciations in the dictionary gives no significant improvement in recognition (Liu and Fung, 2003b). To model acoustic confusions, we treat these accent-specific units as hidden models and adjust the mixture distributions of the pretrained baseform models through the use of mixture components from the hidden models by acoustic model reconstruction. The acoustic model reconstruction is equivalent to tree merging in the decision tree based triphone model structure. This approach aims at refining the pretrained baseform models to achieve a high discriminative ability for the high degree of acoustic confusions, while keeping the model robustness to cover the flexible acoustic variations in accented speech.

1. Auxiliary decision trees for accent-specific triphone units

Context-dependent triphone models are commonly used in current ASR systems for high recognition performance. To limit the model complexity and reduce redundant Gaussian components, decision tree based state tying approach is commonly used (Young, 1999; Hwang *et al.*, 1996). Decision trees for accent-specific triphone units are called *auxiliary decision trees*, compared to *standard decision trees* of baseform triphones. In our system, the structure of triphones for accent-specific units differs from that of baseform triphones only in terms of the central unit. The central unit in an accent-specific tree is a baseform to surface form pair (e.g.

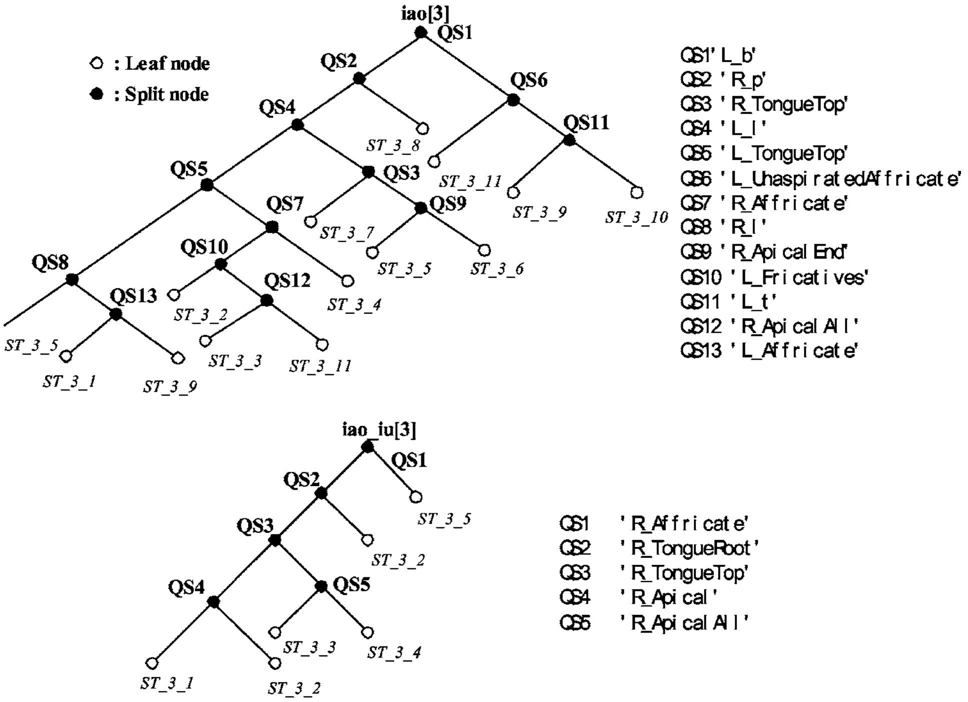

FIG. 7. The auxiliary decision tree of "iao_iu[3]" vs the standard decision tree of "iao[3]".

"*iao_iu*"). Compared to standard decision trees, auxiliary decision 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 auxiliary trees is enlarged to include accent-specific units. The tree size is smaller than that of standard decision trees due to the small training sample of phone units with high acoustic confusions.

The topology of the auxiliary decision trees represents accent variation characteristics. Figure 7 shows an auxiliary tree of "iao_iu" and a standard tree of "iao" at the final, state three. Nearly all the questions for tree splitting of the auxiliary decision tree are right-dependent phonetic questions, while the standard decision tree has both the right-dependent and left-dependent questions. This means that a lot of acoustic variations from /iao/ to /iu/ occur at the final end of the pronunciation. Right-context information is therefore more important than left-context information for accent-specific triphone unit "iao_iu." This is probably because Cantonese speakers tend to move /iao/ to /iu/ at the end of the phone owing to the ingrained influence of their native language.

2. Acoustic model reconstruction through decision tree merge

Auxiliary decision trees representing accented speech and standard decision trees representing standard speech are merged for better recognition of both accented and standard speech.

We merge the leaf nodes of auxiliary decision trees into the related nodes of the standard tree for acoustic model reconstruction as shown in Fig. 8. Through decision tree merge, the pretrained acoustic models are reconstructed to include Gaussian mixture distributions from accent-specific triphone models. As a result, the structure of the Gaussian distribution is adjusted and more Gaussians borrowed from tied states of auxiliary decision trees may locate at the dis-

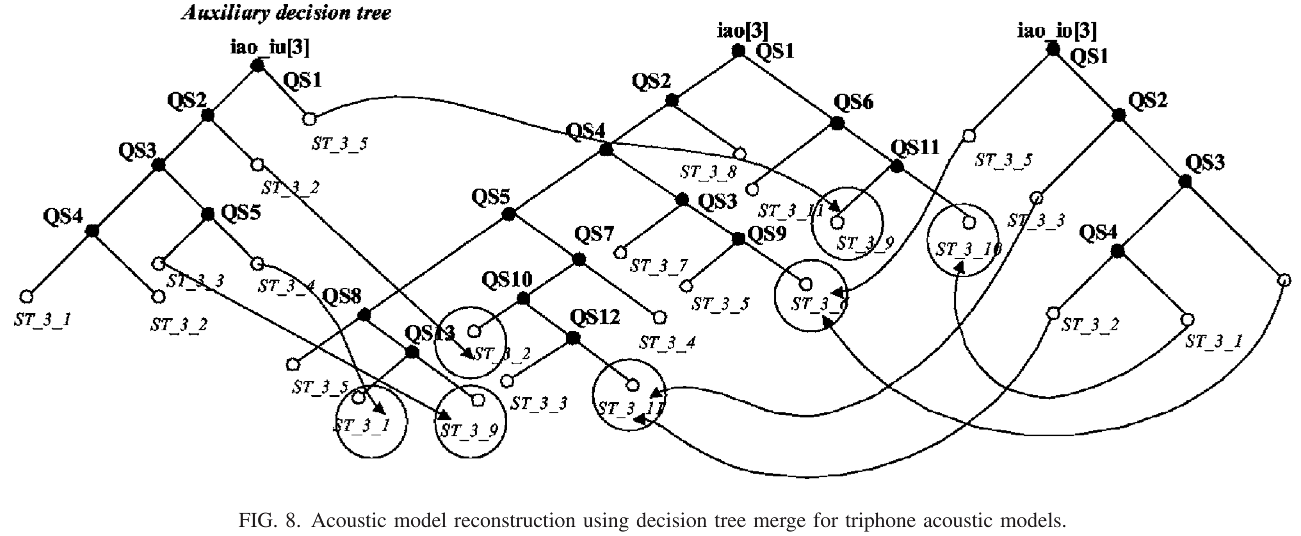

FIG. 8. Acoustic model reconstruction using decision tree merge for triphone acoustic models.

tribution boundaries to cover the variant pronunciations within the accented speech. We use acoustic distance measure of Eq. (5) to determine the mapping relation of tied states between auxiliary decision trees and their related standard trees.

Figure 8 also shows that not all leaf nodes of standard decision tree are mapped to those of auxiliary decision trees. Some nodes have more than one mapping nodes while some nodes have none. The number of mapping nodes is determined by the coverage of the original pretrained model and training samples. For example, the leaf node “ST_ 3_11” of standard decision tree includes mapping nodes from two different auxiliary decision trees in order to model the accented pronunciation changes from /iao/ to /iu/ and from /iao/ to /ao/, while leaf node “ST_3_5” has no mapping node.

According to Fig. 8, the new output distribution of the reconstructed model is represented as

$$P'(x|b) = \lambda P(x|b) + (1-\lambda)\sum_{i=1}^{N} P(x|s_i)P(s_i|b), \qquad (8)$$

where $P(x|b)$ is the output distribution of the pretrained baseform model, λ is a linear interpolation coefficient for combining different acoustic models. The coefficient is the probability of the baseform model being recognized as itself. For instance, if “p[2],”, i.e., the second state of the baseform unit /p/, has 70% probability to be recognized as “p[2]” and 30% probability as other alternate surface forms from the training data, then $\lambda = 0.7$. In addition, $i = 1, 2, \ldots, N$, and N is the total number of merged nodes from auxiliary decision trees; s_i is one possible surface form state from auxiliary decision trees with respect to the baseform state. If a certain leaf node of standard decision tree has no mapping modes, then $N = 0$ and $\lambda = 1$. $P(s_i|b)$ is the confusion probability between the accent-specific unit model and baseform model, which can be estimated from confusion matrix or from state-level pronunciation modeling (Liu and Fung, 2003a; Saraclar, Nock, and Khudanpur, 2000).

V. RECOGNITION EXPERIMENTS

A. Experimental setup

We evaluate our algorithms in a Chinese telephony short phrase recognition task. All speech data were sampled at 8 kHz and 8 bit-rate. The baseform acoustic model was trained using 100 speakers’ utterances with around 50 h of native Mandarin speech. two-thousand continuous utterances with 23 685 syllables from 20 Cantonese-accented speakers (DATA1) were used to extract the accent-specific units. The HMM topology is three-states, left-to-right without skips, and continuous. The acoustic features are 13MFCC, 13ΔMFCC and 13ΔΔMFCC. Twentyone standard initials and 38 finals were used to generate context-independent HMMs. We used the HTK decision tree based state tying procedures to build 12 Gaussian-component triphone models with 5500 tied states. The test data consist of two parts: the first test set (Test_set1) includes 9 speakers (4 females and 5 males) 900 Cantonese-accented utterances apart from DATA1; the second test set (Test_set2) consists of 900 standard Mandarin utterances selected from 9 native speakers (4 females and 5 males), and is used for performance comparison. In order to evaluate the recognition performance gains solely from phonetic and acoustic modeling, free from other high level information, all the utterances of the test sets are Chinese short phrases without word n-grams.

TABLE II. A comparison of WER of using multiple pronunciation dictionaries based on accent-specific units compared to using conventional reweighed and augmented dictionary.

	Word error rate (WER)%	
System	(Test_set1) Accented speech	(Test_set2) Mandarin speech
Baseline	20%	7.9%
Multiple pronunciation dictionary (Dict1)	17.3%(−2.7)	8.1%(+0.2)
Reweighted and augmented dictionary (Dict2)	18%(−2.0)	7.7%(−0.2)
Selected multiple pronunciation dictionary (Dict3)	**16.9%(−3.1)**	**7.7%(−0.2)**

B. Modeling accent-specific units with high phonetic confusion and low acoustic confusion

Using DATA1 as the development set, we obtained 79 accent-specific units with high phonetic confusions. We first used these units to generate a multiple pronunciation dictionary (Dict1) and compared its performance with respected to a conventional reweighed and augmented dictionary (Dict2) that is based on minimum count and minimum out relative frequency criteria (Byrne *et al.*, 2001; Huang *et al.*, 2000; Liu *et al.*, 2000). The results are shown in Table II. We can see that augmenting a multiple pronunciation dictionary with these high phonetic confusion units gives us an encouraging 2.7% absolute WER reduction compared to the baseline and a slight 0.7% absolute WER reduction with respect to using Dict2.

Furthermore, we compared the tendency of initial/final error rate (IFER) to that of word error rate (WER) by varying the extended phone unit numbers. As shown in Fig. 9, we found that lower IFER does not always lead to lower WER. In an extreme case, introducing more accent-specific units leads to the degradation of recognition performance. We believe that the inability of transferring the lower IFER to lower SER is caused by lexical confusion. The accent-specific high phonetic confusion units in Dict2 include both high acoustic confusions as well as low acoustic confusion. This shows that we need to model these two classes of phonetic units separately. Using the asymmetric acoustic distance measure, only 22 units with low acoustic confusions are selected to form alternative pronunciations and generated a selected multiple pronunciation dictionary (Dict3). Table II shows that using Dict3 is more efficient to cover phonetic confusions in accented speech than using Dict1 and Dict2, yielding additional 0.4% and 1.1% absolute WER reductions, respectively.

Moreover, we can see that the use of Dict3 on native Mandarin speech (test_set2) does not lead to any perfor-

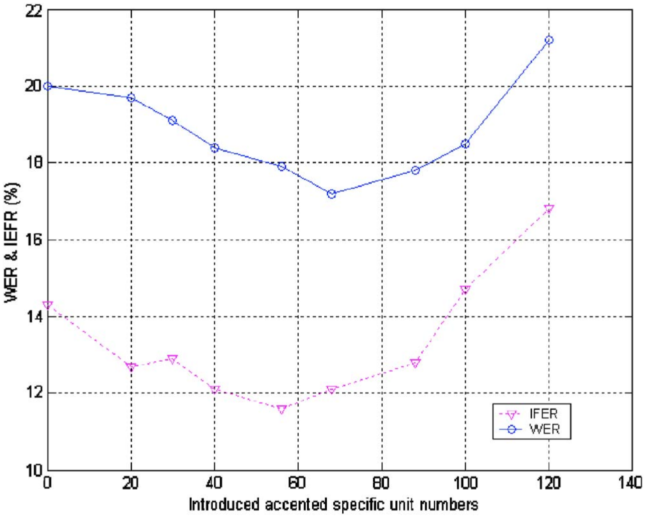

FIG. 9. WER and IFER with different amount of selected accent-specific units. Lower IFER does not always lead to lower WER.

mance degradation since there are no acoustic and lexical confusions between the additional pronunciations and the originally canonical pronunciations. On the other hand, using the conventional augmented dictionary, Dict1, with accent-specific units leads to worse performance on test_set2. That is, additional acoustic and lexical confusions are introduced when alternative pronunciations related to accent effects are added into the dictionary. These results support our claim that adding acoustically confusable phone units in an augmented dictionary leads to more decoder error in recognition.

C. Modeling accent-specific units with high phonetic confusions and high acoustic confusions

Fifty-seven units from the original 79 accent-specific units were extracted as units with high phonetic and acoustic confusions. We constructed 171 auxiliary decision trees with 967 tied states for these 57 accent-specific triphone units. Through acoustic model reconstruction, 967 tied states were merged into the pretrained 5500 tied states of 177 standard decision trees. The reconstructed model included 77 604 Gaussian components and each state has 14.1 Gaussians on average. To make a fair comparison, we generated an enhanced baseform model with 5500 tied states and 14 Gaussian-component per state. The recognition performances are shown in Table III.

TABLE III. Our approach outperforms MAP adaptation, enhanced acoustic model, and augmented dictionary.

	Word error rate (WER) %	
System	(Test_set1) Accented speech	(Test_set2) Standard speech
Baseline	20%	7.9%
Enhanced HMMs with 14 Gaussians per state	18.6% (−1.4)	7.5% (−0.4)
Baseline HMM with MAP adaptation using accented data	15.1% (−4.9)	15.7% (+6.8)
Reconstructed HMMs with selected accent-specific units	**15.2% (−4.8)**	**7.1% (−0.8)**

We can see that using the reconstructed acoustic model gives a significant 4.8% absolute WER reduction compared to the baseline, and an additional 3.4% reduction with respect to using enhanced HMM at the same model complexity. The reason lies in the fact that the mixture distribution of our reconstructed model includes borrowed Gaussians from accent-specific unit models, and adjusts the structure of the original mixture distribution and enables more Gaussians at the mixture boundaries to cover the acoustic confusions in accented speech. On the other hand, directly increasing Gaussian components in the enhanced model results in poor estimation of some Gaussians with available training data. Meanwhile, most of the increased Gaussians may converge around the global mean to handle the majority of pronunciation with small variations, and there are not sufficient Gaussians at the boundary of mixture distributions.

One advantage of using reconstructed acoustic models is that our method provides significant improvement for accented speech task without sacrificing the performance on native Mandarin speech. In comparison, the use of MAP adaptation approach gives a good 4.9%WER reduction on accented speech, while leading to a serious performance degradation (6.8% WER increase) on native Mandarin speech. Through MAP adaptation, the parameters of acoustic model are adjusted to handle accented speech and are no longer suitable for native speech. However, our reconstructed model includes its own Gaussians from pretrained acoustic model as well as those borrowed from accent-specific unit models. The borrowed Gaussians are used only to adjust the structure of original mixture distribution and not to change parameters. These two Gaussian distributions cover the acoustic samples either with small deviation in native speech or with high deviation in accented speech.

In addition, Gaussian mixture sharing and clustering across phonetic models with minimal average distortion have been shown to be efficient in improving model robustness for acoustic confusions (Huang and Jack, 1989; Nakamura, 2002). The question is whether the same amount of WER reduction can be achieved by straightforward Gaussian mixture sharing. To answer this question, a comparison of recognition performance between our acoustic model reconstruction with selected accented-specific units and Gaussian mixture sharing of baseline model is illustrated in Table IV. Note that in the decision tree-bases state tying triphone models with Gaussian sharing, based on the extended accent-specific units, an additional 7683 mixture weights are added as new parameters.

In addition, our reconstructed model gives 3% absolute WER reduction in relation to Gaussian sharing models on accented speech. In Gaussian mixture sharing, only the shared parameters are trained efficiently, while the shared Gaussians may not cover the acoustic confusions that locate at the boundary of mixture distributions. On the other hand, our reconstructed model includes more Gaussian components borrowed from accent-specific unit models at the boundary of the mixture distributions, when the confusing

TABLE IV. Our approach outperforms the baseline, and modeling phonetic or acoustic confusion alone.

	Word error rate (WER) %	
System	(Test_set1) Accented speech	(Test_set2) Standard speech
Baseline	20%	7.9%
Baseline model with Gaussian mixture sharing	18.2%(−1.8)	7.0%(−0.9)
Model trained using selective surface form tran- scriptions (modeling phonetic confusion only)	19.1%(−0.9)	7.6%(−0.3)
Reconstructed HMMs with selected accent-specific units (modeling acoustic confusion only)	15.2%(−4.8)	7.1%(−0.8)
Reconstructed HMMs and selected multiple pronunciation dictionary (our approach)	**14.3%(−5.7)**	**7.1%(−0.8)**

acoustic samples fall into this mixture distribution, a higher acoustic likelihood score is obtained compared to using Gaussian sharing models.

It was shown in Riley *et al.* (1999) that acoustic models can be trained by using the surface form transcriptions iteratively. We compare this approach with our reconstructed model in Table IV. It has been reported in Riley *et al.* (1999) and also shown here that their method gives limited performance improvement. We note that the selection of surface form transcriptions is mainly based on *phonetic confusions* not acoustic confusions. On the other hand, the recognition is primarily based on the acoustic distance, not the phonetic distance, so WER will not be reduced if the acoustic distance among the units remains unchanged. We give an example in Fig. 10 of the acoustic distance of /zh/ with respect to other Chinese initials/finals in baseline model and retrained model using selective surface form transcriptions and show that there is no distinct acoustic distance. This also explains why very limited improvement was achieved by using retrained acoustic models based on selective phone-level transcriptions in Riley *et al.* (1999). The evidence again indicates that phonetic and acoustic confusions should be treated separately in accented speech recognition. Moreover, using selected multiple dictionary as well as acoustic model reconstruction provides a significant 5.7% WER reduction without sacrificing the performance on native Mandarin speech. That is, our approach can be applied to a single system for both accented and native speech recognition.

Last but not the least, we show the performance comparison between our approach, the baseline approach, methods using only phonetic confusion modeling and a method using only acoustic confusion in Table V and show that modeling phonetic and acoustic confusable units separately gives the best performance on accented speech as well as standard speech recognition.

VI. CONCLUSIONS

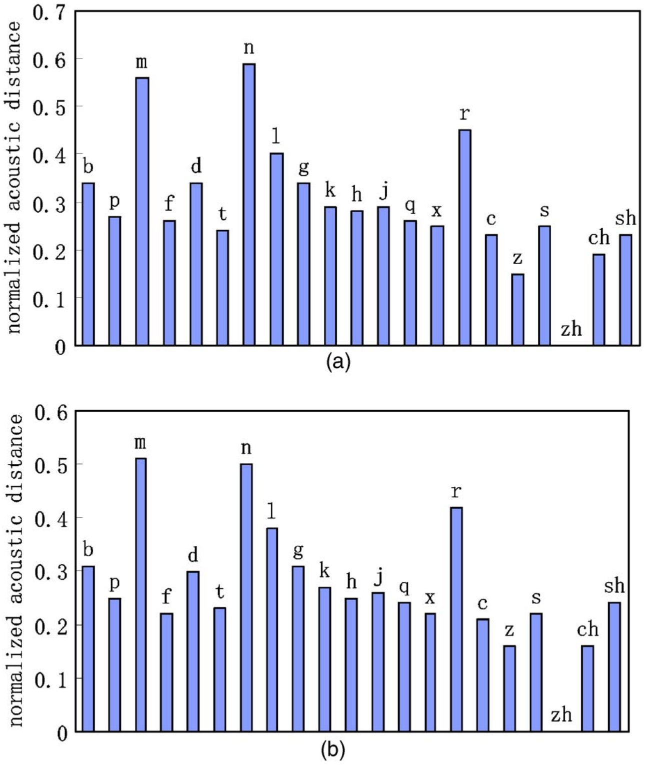

FIG. 10. The normalized acoustic distance of "zh" in relation to other Chinese initials in baseline model and surface form retrained model.

We study the effects of phonetic confusions and acoustic confusions in accented speech. We suggest that phonetic and acoustic confusions are different yet correlated in accented speech. We suggest that only phone units which lead to high phonetic confusions in accented speech cause recognition errors. Among these units, there are those that also have high acoustic confusions and others with low acoustic confusions. We propose to model these two classes of phone units differently for better recognition performance on both accented and standard speech. We use likelihood ratio test to select units with high phonetic confusions and we propose an asymmetric acoustic distance measure to describe the unidirectional properties of acoustic confusions in accented speech. In addition, we separated accent-specific confusions from data and recognizer-related confusions using distance measure and pronunciation phonological rules.

We propose incorporating only those accent-specific phonetic units with low acoustic confusions in a multiple pronunciation dictionary in order to reduce phonetic confusions and avoid lexical confusion at the same time. Mean-

TABLE V. Our approach outperforms the baseline, and modeling phonetic or acoustic confusion alone.

	Word error rate (WER) %	
System	(Test_set1) Accented speech	(Test_set2) Standard speech
Baseline	20%	7.9%
Multiple pronunciation dictionary (modeling phonetic confusion alone)	17.3%(−2.7)	8.1%(+0.2)
Reconstructed HMMs (modeling acoustic confusion alone)	15.2%(−4.8)	7.1%(−0.8)
Reconstructed HMMs and selected multiple pronunciation dictionary	**14.3%(−5.7)**	**7.1%(−0.8)**

while, for accent-specific units with high acoustic confusion, we propose using decision tree merging with acoustic model reconstruction to achieve a high discriminative ability for reducing acoustic confusions within phonetic unit models. This approach aims at using the selected accent-specific units as hidden models to adjust the structure of mixture distributions of standard speech baseform models to cover more acoustic variability so as to model acoustic confusions in accented speech.

Experimental results on Cantonese-accented Mandarin speech show that using the selected multiple pronunciation dictionary to model phonetic confusions provides WER reductions of 3.1% and 1.1% in absolute terms, compared to baseline and using conventional reweighted and multiple pronunciation dictionary. Through the use of acoustic model reconstruction, we achieve a significant 4.8% absolute WER reduction for accented speech compared to 1.4% using increasing Gaussian components and 1.8% by Gaussian mixture sharing. The combination of modeling phonetic confusions and acoustic confusions yields a 5.7% reduction. Compared to using MAP adaptation, our method provides a better WER reduction on accented speech recognition without sacrificing the performance on native, standard speech. Our approach can be applied to a single system to handle both accented and standard speech, and even speech with multiple accents.

ACKNOWLEDGMENTS

This work is partially supported by CERG No. HKUST6206/03E of the Hong Kong Research Grants Council, and DAG No. 03/04.EG30 of HKUST.

[1] For example, the acoustic distance from “zh” to “z” is entirely different from that of “z” to “zh” in Cantonese-accented Mandarin speech. The distance from “zh” to “z” is much smaller than that of “z” to “zh.”

Bacchiani, M., and Ostendorf, M. (**1998**). “Joint acoustic unit design and lexicon generation,” *Proceedings of the Workshop on modeling pronunciation variation for ASR, 1998*, 7–12.

Byrne, W., Venkataramani, V., Kamm, T., Zheng, F., Fung, P., Liu, Y., and Ruhi, U. (**2001**). “Automatic generation of pronunciation lexicons for Mandarin spontaneous speech,” *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT.

Chen, Y. J., Wu, C. H., Chiu, Y. H., and Liao, H. C. (**2002**). “Generation of robust phonetic set and decision tree for Mandarin using chi-square testing,” Speech Commun. **38**, 349–364.

Chou, W., Juang, B. H., and Lee, C. H. (**1992**). “Segmental GPD training of HMM based speech recognizer,” *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, San Francisco, CA, pp. 473–476.

Cover, T. M., and Thomas, J. A. (**1991**). *Elements of Information Theory* (Wiley, New York).

Fung, P., Byrne, W., Zheng, F., Kamm, T., Liu, Y., Song, Z., Venkataramani, V. and Ruhi, U. (**2000**). “Pronunciation modeling of Mandarin casual speech,” Final Report, The Johns Hopkins University Summer Workshop.

Holter, T. (**1997**). “Maximum likelihood modeling of pronunciation in automatic speech recognition,” Ph.D. thesis, the Norwegian University of Science and Technology.

Huang, Ch. *et al.* (**2000**). “Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition,” *Proccedings of the International Conference on Spoken Language Processing (ICSLP2000)*, Beijing, China.

Huang, J. H. (**1987**). *Chinese Dialects* (Xia Men University Press, Xia Men, China) (Chinese version).

Huang, X., and Jack, M. (**1989**). “Unified techniques for vector quantization and hidden Markov modeling using semi-continuous models,” *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, Glasgow, Scotland, pp. 639–642.

Hwang, M. Y. (**1993**). “Subphonetic acoustic modeling for speaker-independent continuous speech recognition,” Ph.D. thesis, Carnegie Mellon University.

Hwang, M. Y., Huang, X. D., and Alleva, F. A. (**1996**). “Predicting unseen triphones with senones,” IEEE Trans. Speech Audio Process. **4**, 412–419.

Juang, B. H., and Katagiri, S. (**1992**). “Discriminative learning for minimum error classification,” IEEE Trans. Signal Process. **40**, 3043–3054.

Jurafsky, D., Ward, W., Zhang, J. P., Herold, K., Yu, X. Y., and Zhang, S. (**2001**). “What kind of pronunciation variation is hard for triphones to model?,” *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, (ICASSP2001)*, Salt Lake City, UT.

Katagiri, S., Juang, B. H., and Lee, C. H. (**1998**). “Pattern recognition using a family of design algorithm based upon the generalized probabilistic descent method,” Proc. IEEE **86**, 2345–2373.

Lee, T., Lau, W., Wong, Y. W., and Ching, P. C. (**2002**). “Using tone information in Cantonese continuous speech recognition,” ACM Transactions on Asian Language Information Processing **1**(1), 83–102.

Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V., and Chen, X. (**2000**). “CASS: A phonetically transcribed corpus of Mandarin spontaneous speech,” *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

Li, X., and King, I. (**1999**). “Gaussian mixture distance for information retrieval,” *Proceedings of the 1999 International Joint Conference on Neural Networks*, Washington DC, pp. 2070–2075.

Liu, M. K., Xu, B., Huang, T., and Li, C. (**2000**). “Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling,” *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, (ICASSP2000)*, Istanbul, Turkey, pp. 1929–1932.

Liu, W. K., and Fung, P. (**1999**). “Fast accent identification and accented speech recognition,” *Proceedings of the IEEE International Conference Acoustics, Speech, Signal Processing (ICASSP)*, Phoenix, A2.

Liu, Y., and Fung, P. (**2001**). “Estimating pronunciation variations from acoustic likelihood score for HMM reconstruction,” *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, pp. 1425–1428.

Liu, Y., and Fung, P. (**2003a**). “Modeling partial pronunciation variations for spontaneous Mandarin speech recognition,” Comput. Speech Lang., **17**, 357–379.

Liu, Y., and Fung, P. (**2003b**). “Partial change accent models for accented Mandarin speech recognition,” *Proceedings of the IEEE Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands.

Liu, Z., and Huang, Q. (**2000**). “A new distance measure for probability distribution function of mixture type,” *Proceedings of the IEEE International Conference. Acoustics, Speech, Signal Processing, (ICASSP2000)*, Istanbul, Turkey, pp. 1345–1348.

Manning, C. D., and Schütze, H. (**1999**). *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).

Nakamura, A. (**2002**). “Restructuring Gaussian mixture density functions in speaker-independent acoustic models,” Speech Commun. **36**, 277–289.

Peters, S., and Stubley, P. (**1998**). “Visualizing speech trajectories,” *Proceedings of ESCA Totorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, Netherlands, 1998, pp. 97–101.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., Mcdonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G. (**1999**). “Stochastic pronunciation modeling from hand-labeled phonetic corpora,” Speech Commun., **29**, 209–224.

Saraclar, M., Nock, H., and Khudanpur, S. (**2000**). “Pronunciation modeling by sharing Gaussian densities across phonetic models,” Comput. Speech Lang., **14**, 137–160.

Sproat, R. (**2001**). “Pmtools: A pronunciation modeling toolkit,” *Proceedings of the Fourth ISCA Tutorial and Research Workshop on Speech Synthesis*, Blair Atholl, Scotland.

Stevens, K. N., *Acoustic Phonetics* (MIT Press, Cambridge, MA, 1998).

Strik, H., and Cucchiarini, C. (**1999**). "Modeling pronunciation variation for ASR: A survey of the literature," Speech Commun. **29**, 225–246.

Tomokiyo, L. M. (**2001**). "Recognizing non-native speech: Characterizing and adapting to non-native usage in LVCSR," Ph.D. thesis, Carnegie Mellon University.

Tsai, M. Y., and Lee, L. S. (**2003**). "Pronunciation variation analysis based on acoustic and phonetic distance measures with application examples on Mandarin Chinese," *Proceedings of the IEEE Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, pp. 117–121.

Young, S. (**1999**). *The HTK Book* (Entropic Cambridge Research Laboratory, Cambridge).