INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 7, 141–153, 2004

Accent Issues in Large Vocabulary Continuous Speech Recognition

CHAO HUANG, TAO CHEN* AND ERIC CHANG
Microsoft Research Asia, 5F, Sigma Center, No. 49, Zhichun Road, Beijing 100080, China
chaoh@microsoft.com
tao.chen@ncl.ac.uk
echang@microsoft.com

Abstract. This paper addresses accent[1] issues in large vocabulary continuous speech recognition. Cross-accent experiments show that the accent problem is very dominant in speech recognition. Analysis based on multivariate statistical tools (principal component analysis and independent component analysis) confirms that accent is one of the key factors in speaker variability. Considering different applications, we proposed two methods for accent adaptation. When a certain amount of adaptation data was available, pronunciation dictionary modeling was adopted to reduce recognition errors caused by pronunciation mistakes. When a large corpus was collected for each accent type, accent-dependent models were trained and a Gaussian mixture model-based accent identification system was developed for model selection. We report experimental results for the two schemes and verify their efficiency in each situation.

Keywords: automatic speech recognition, speaker variability, pronunciation modeling, accent adaptation, accent identification

1. Introduction

In recent years, automatic speech recognition (ASR) systems, even in the domain of large vocabulary continuous ASR, have achieved great improvements. There are several commercial systems on the shelves like ViaVoice of IBM, SAPI of Microsoft and NaturallySpeaking of Dragon.

At the same time, speaker variability still affects the performance of ASR systems greatly. Among the factors attributing variability, gender and accent are the most important (Huang et al., 2001). The former has been managed by gender-dependent models. However, there is relatively little research on accented speech recognition, especially for speakers who have the same mother tongue, but varying regional accents caused by the dialects of the speakers.

There are two speech research areas related to accent issues: accent adaptation through pronunciation modeling and accent identification. It is known that speakers with heavy accents tend to make more pronunciation errors in terms of the standard pronunciation. Experimental analysis (Huang et al., 2000) showed that this type of errors constituted a considerable proportion of total errors. In addition, it was observed that speakers from the same accent regions had similar tendencies in mispronunciations. Based on the above facts, pronunciation modeling emerged as a solution. The basic idea was to catch typical pronunciation variations through a small amount of data and encode them into a so-called accent-specific dictionary. Conventional pronunciation modeling methods were categorized by two criteria (Strik and Cucchiarini, 1998): data-driven vs knowledge-driven and formalized information representation vs enumerated one. It was also observed that simply adding several alternative pronunciations to the dictionary

*T. Chen participated in the work from 2001 to 2002 as an intern at Microsoft Research Asia and is currently with the Centre for Process Analytics and Control Technology, University of Newcastle upon Tyne, NE1 7RU, U.K.

may increase the confusability of words (Riley et al., 1999).

In accent identification, current research focuses on classifying non-native accents. In addition, most systems (Hansen and Arslan, 1995; Teixeira et al., 1996; Fung and Liu, 1999) were built on hidden Markov models (HMM). HMM training is time-consuming. Furthermore, HMM training is a supervised procedure and transcriptions are needed. The transcriptions are either labeled manually, or obtained from a speaker-independent model in which the alignment errors will certainly degrade the identification performance.

In this paper, accent issues are addressed in a general framework. The impact of accented speech on recognition performance was explored. We trained a model for each accent and collected test data from different accents. Cross-accent speech recognition experiments showed that error rate increased up to 40–50% when the acoustic model and the test data were from different accents. Then principal component analysis (PCA) and independent component analysis (ICA) were used to investigate dominant factors in speaker variability. Experiments confirmed qualitatively the fact that the accent problem is very crucial in speech technologies.

To deal with accent variability, we suggested two solutions according to different applications. When only a speaker-independent model and some amount of adaptation data from an accent group were available, a Pronunciation Dictionary Adaptation (PDA) was developed to reduce error rate caused by mis-pronunciation. We extended the syllable-based context (Liu et al., 2000) to be more flexible: context level was decided by the amount of data for PDA. In addition, some previous work (Riley and Ljolje, 1996; Humphries and Woodland, 1998; Liu et al., 2000) utilized pronunciation variation information to re-score the N-best hypothesis or lattices resulting from the baseline system. However we developed a one-pass search strategy to unify all the information from acoustic, language and accent models.

When a large amount of training data for each accent was available, we built Accent-Dependent (AD) models similar to gender-dependent ones. Although it may be not efficient to provide multiple models in desktop applications, it is still practical in a client-server framework. The core problem of such a strategy is to select the proper model for each test speaker automatically. We propose a Gaussian Mixture Model (GMM)-based accent identification method, whose training process is unsupervised. After identification, the most likely accent dependent model is selected for recognition.

Although all our experiments were conducted on a Mandarin ASR system, the investigations and proposed adaptation methods can be applied to other languages.

This paper is organized as follows. Section 2 investigates the accent problem from two aspects: quantitatively and qualitatively or cross-accent speech recognition experiments and a high-level analysis by PCA and ICA. In Section 3 we propose a pronunciation dictionary adaptation to reduce error rate caused by mispronunciation. In Section 4 we describe an automatic accent identification method based on a Gaussian mixture model and verify its effectiveness in selecting an accent dependent model. Finally conclusions were given in Section 5.

2. Impact of Accent on Speech

As described in Section 1, accent is one of the challenges in current ASR systems. Before introducing our solutions to accent issues, we investigated the impact of accent on speech from two views. First, cross-accent speech recognition experiments were carried out. Second, multivariate analysis tools, PCA and ICA, were applied to confirm quantitatively the importance of accent in speaker variability.

2.1. Cross-Accent Speech Recognition Experiments

In order to investigate the impact of accent on state-of-the-art ASR systems, extensive experiments were carried on the Microsoft Mandarin speech engine (Chang et al., 2000), which has been successfully released with Office XP and SAPI. In the system, tone-related information, which is very helpful in ASR for tonal languages, also was integrated through pitch features and tone modeling. All the speech recognition experiments in this paper were based on this solid and powerful baseline system.

The training corpora and model configurations are listed in Table 1. Three typical Mandarin accents, Beijing (BJ), Shanghai (SH) and Guangdong (GD) were considered. For comparison, an accent-independent model (X6) was also trained based on $\sim$3000 speakers. In addition, gender-dependent models were trained and used in all experiments. Table 2 lists the test corpora.

Table 3 shows the recognition results. Character Error Rate (CER) was used for evaluation. It is easily

Table 1. Summary of training corpora for cross-accent experiments.

Model tag	Training corpus configurations	Approx. amount of data
BJ	1500 Beijing Speakers	330 hours
SH	1000 Shanghai Speakers	220 hours
GD	500 Guangdong Speakers	110 hours
X6	BJ + SH + GD (3000 Speakers)	660 hours

concluded that accent variations between the training and test corpus degrade recognition accuracy significantly. Compared with an accent-dependent model, cross-accent models increased error rate up to 40–50% while an accent-independent model (X6) increased error rate by 15–30%. It should be noted that the great performance difference on three testing sets given the same acoustic model were due to the different complexities of the sets, shown as character perplexity (PPC^2) in Table 2.

2.2. *Investigation of Accent Variability*

In this subsection, we investigate some of the key factors in speaker variability. What these factors are and how they correlate with each other are of great concern in speech research. One of the difficulties in this investigation was the complexity of the speech model. There usually are a huge number of free parameters associated with a set of models. Thus, the representation of a speaker is usually high-dimensional when different phones are taken into account.

Fortunately, several powerful tools, such as principal component analysis (PCA) (Hotellings, 1933) and independent component analysis (ICA) (Hyvarinen and Oja, 2000), are available for high dimension multivariate statistical analysis. They have been applied successfully in speech analysis (Malayath et al., 1997; Hu, 1999). PCA decorrelates second order moments corresponding to low frequency properties and extracts orthogonal principal components of variations. ICA, while not necessarily orthogonal, makes unknown linear mixtures of multi-dimensional random variables as statistically independent as possible. It not only decorrelates the second order statistics but also reduces higher-order statistical dependency. ICA representation manages to capture the essential structure in the data of many applications including feature extraction and blind source separation (Hyvarinen and Oja, 2000).

In this subsection, we present a subspace analysis method for the analysis of speaker variability. A transformation matrix obtained from maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) was adopted as the original representation of the speaker characteristics. Generally, each speaker was represented by a super-vector that included different regression classes, with each class a vector. Important components in low-dimensional space were extracted by PCA or ICA. It is hypothesized that the dominant components extracted by PCA or ICA represent the key factors of speaker variability. More details of this method can be found in Huang et al. (2001).

2.2.1. Speaker Representation. A speaker adaptation model (MLLR transformation matrix) was adopted to represent the characteristics of a speaker. Such a representation provides a flexible way to control the model parameters according to the available adaptation data. To reflect a speaker in detail, up to 65 regression classes are used in accordance with Mandarin phonetic structure. Limited by adaptation data, only six single vowels (/a/, /i/, /o/, /e/, /u/, /v/) were selected empirically as supporting regression classes.[3] Also, experiments (Huang et al., 2001) showed that using only offset vectors in MLLR can achieve better results in gender classification. In the end, some acoustic features are pruned

Table 2. Summary of test corpora for cross-accent experiments (PPC is perplexity of character according to the language model developed on a 54 k dictionary).

Test sets	Gender	Accent	Speakers	Utterances	Characters	PPC
BJ-M	Male	Beijing	25	500	9570	33.7
BJ-F	Female	Beijing	25	500	9423	
SH-M	Male	Shanghai	10	200	3243	59.1
SH-F	Female	Shanghai	10	200	3287	
GD-M	Male	Guangdong	10	200	3233	55–60
GD-F	Female	Guangdong	10	200	3294	

Table 3. Character error rate (%) for cross-accent experiments.

	Different accent test sets		
Model	BJ	SH	GD
BJ	**8.81**	21.85	31.92
SH	10.61	**15.64**	28.44
GD	12.94	18.71	**21.75**
X6	9.02	17.59	27.95

by experiments to eliminate poorly estimated parameters. In summary, after MLLR adaptation, the following strategy was adopted to represent a speaker.

- Supporting regression classes: six single vowels (/a/, /i/, /o/, /e/, /u/, /v/).
- Offset items in MLLR transformation matrices.
- 26 dimensions of acoustic features (13-d MFCC + ΔMFCC).

As a result, a speaker is typically described by a supervector of $6*1*26 = 156$ dimensions before PCA/ICA projection.

2.2.2. Experiments. The whole corpus consisted of 980 speakers, with 200 utterances per speaker. Speakers are from two accent areas: Beijing (BJ) and Shanghai (SH). The gender and accent distributions are summarized in Table 4.

All the speakers are concatenated into a 980×156 matrix. Then speakers are projected onto the top six components extracted by PCA and a new whitened matrix of 980×6 is obtained. The matrix is fed to ICA (implemented according to the FastICA algorithm proposed by Hyvarinen and Oja (2000)). Figures 1 and 2 show the projections of all the speakers onto the first two independent components. The horizontal axis is the speaker index whose alignment is: BJ-F (1–250), SH-F (251–440), BJ-M (441–690) and SH-M (691–980).

It can be concluded from Fig. 1 that the first independent component corresponds to gender characteristics of a speaker: projections on this component almost separate all speakers into gender categories. In Fig. 2, four subsets occupy four blocks. The first and the third one correspond to the Beijing accent while the second and the fourth one correspond to Shanghai. It is obvious that this component has strong correlations with accent. A 2-d illustration of an ICA projection is shown in Fig. 3. It can be concluded that accent and gender are the main components that constitute the speaker space.

Table 4. Speaker distribution for speaker variability analysis.

	Beijing	Shanghai
Female	250 (BJ-F)	190 (SH-F)
Male	250 (BJ-M)	290 (SH-M)

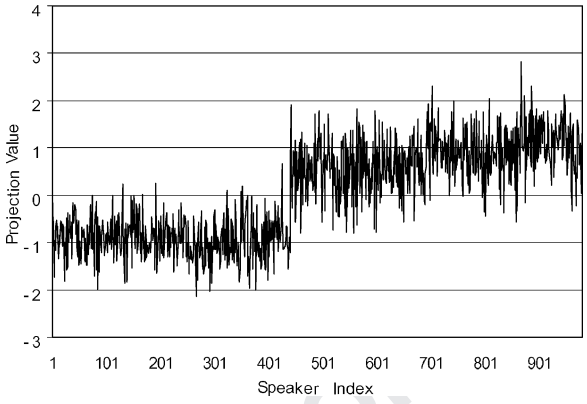

Figure 1. Projection of all the speakers onto the first independent component (The first block corresponds to the speaker sets BJ-F and SH-F, and the second block corresponds to the sets BJ-M and SH-M).

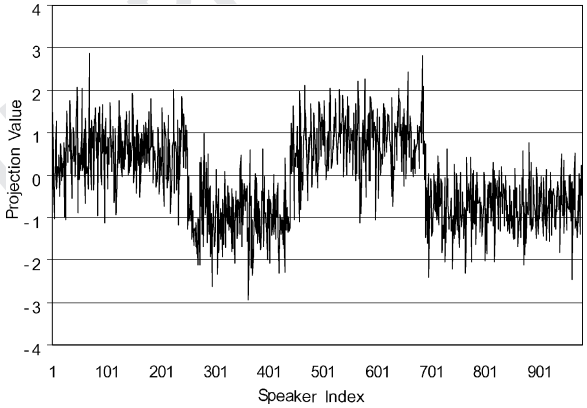

Figure 2. Projections of all speakers onto the second independent component. (The four blocks correspond to the speaker sets BJ-F, SH-F, BJ-M and SH-M, from left to right).

2.3. Summary

In this section, both cross-accent experiments and speaker variability analysis showed that accent is one of the most important factors leading to fluctuating performance of an ASR system. The accent problem is very crucial, especially in countries with large areas. Across

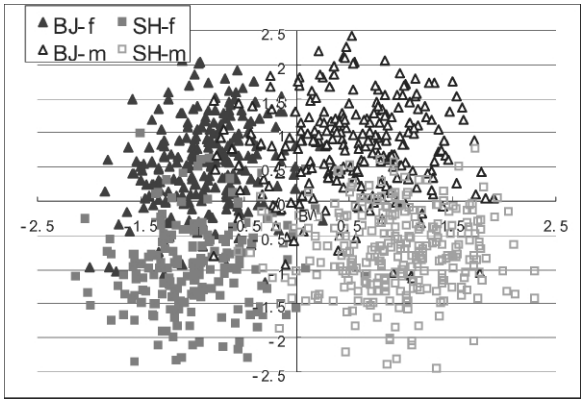

Figure 3. Projection of all the speakers onto the space constructed by the first (horizontal axis) and the second (vertical axis) independent components.

China, almost every province has its own dialect.[4] When speaking Mandarin, a person's dialect usually brings a heavy accent to his/her speech. Different solutions can be developed according to applications. When some amount of data are available, adaptation methods, such as MLLR and MAP (Lee et al., 1991), can be used to reduce the variations between the baseline model and the test speaker. In the following section, an adaptation method based on a pronunciation dictionary is described that it decreases recognition errors caused by the speaker's own pronunciation mistakes. As a pronunciation model is complementary to a acoustic model, this method is expected to achieve more improvement over the baseline system when combined with standard MLLR adaptation. In other situations, a large amount of speech data may be collected from each accent region while only several utterances are available for a test speaker. We trained one model for each accent, and a distance criterion was developed to select an accent-dependent model according to the limited data. These methods will be discussed in the Sections 3 and 4, respectively.

3. Pronunciation Dictionary Adaptation

From cross-accent experiments in Section 2.1, we found that a speech recognizer built for a certain accent type usually obtains a much higher error rate when applied to speech with another accent. The errors come from two sources. One is misrecognition of confusable sounds by the recognizer. The other one is the speaker's own pronunciation mistakes in terms of standard pronunciation. For example, some Chinese people are not able to differentiate between /zh/ and /z/ in standard Mandarin. Error analysis shows that the second type of errors constitutes a large proportion of the total errors in the cross-accent scenario. Furthermore it is observed that speakers belonging to the same accent region have similar tendencies in mispronunciation.

Based on the above fact, an accent modeling technology named pronunciation dictionary adaptation (PDA) is proposed. The basic idea is to catch the typical pronunciation variations for a certain accent through a small amount of adaptation data and encode these differences into the dictionary (accent-dependent dictionary). Depending on the amount of adaptation data, a dynamic dictionary construction process can be presented in multiple levels such as phoneme, base syllable or tonal syllable. Both context-dependent and context-independent pronunciation models were considered. To ensure that the confusion matrices reflected the accent characteristics, both the occurrences of reference observations and the probability of pronunciation variation were taken into account when deciding which transformation pairs should be encoded. In addition, as pronunciation variations and acoustic deviations are complementary, PDA combined with standard MLLR adaptation was also applied.

Compared with the method proposed by Humphries and Woodland (1998), which synthesizes the dictionary completely from the adaptation corpus, we enhanced the process by incorporating obvious pronunciation variations into the accent-dependent dictionary with varying weights. As a result, the adaptation corpus for catching the accent characteristics could be comparatively small. Essentially, the entries in the adapted dictionary consisted of multiple pronunciations with prior probabilities that reflected accent variation. We extended syllable-based context (Liu et al., 2000) to phone-level and phone-class level, which was decided by the amount of data for PDA. This flexible method can extract the essential variation in continuous speech from a limited corpus while maintaining a detailed description of the effect of articulation on pronunciation variation. Furthermore, tone changes, as a part of pronunciation variation, also can be modeled.

Instead of using pronunciation variation information to re-score the N-best hypothesis or lattices, we developed a one-pass search strategy that unified all kinds of information like acoustic model, language model and accent model about pronunciation variation, in accordance with the existing baseline system.

3.1. Accent Modeling with PDA

Conventional acoustic model adaptation technologies assume that speakers pronounce words in a predefined and unified manner, which is not always valid for accented speech. For example, a Chinese speaker from Shanghai probably utters syllable[5] /shi/ as /si/ in the canonical dictionary. Therefore, a recognizer trained on the pronunciation criterion of standard Mandarin cannot accurately recognize speech from a Shanghai speaker. Fortunately, pronunciation variation between accent groups usually presents certain clear and fixed tendencies. There exist some distinct transformation pairs at the level of phones or syllables. These provide the premise by which to carry out accent modeling through PDA, which can be divided into the following stages.

The first stage is to transcribe available accented speech data by a recognizer based on a canonical pronunciation dictionary. To reflect true pronunciation deviation, no language model was used here. The obtained transcriptions were aligned with the reference ones through dynamic programming. Then error pairs were identified. Only substitution errors were considered. Mapping pairs with few observations or low transformation probabilities were pruned to eliminate those caused by recognition error. For example, as the pair "/si/ -> /ci/" appeared only several times in the corpus, it was regarded as coming from recognition error, not from pronunciation deviation. According to the amount of accented corpus, context-dependent or context-independent mapping pairs with different transformation probabilities were selectively extracted at the level of sub-syllable, base-syllable or tone-syllable.

The second stage is to construct a new dictionary that reflects accent characteristics based on the transformation pairs. We encoded these pronunciation transformation pairs into the original canonical lexicon, and construct a new dictionary adapted to a certain accent. In fact, pronunciation variation was implemented through multiple pronunciations with corresponding weights. All the pronunciation variations' weights corresponding to the same word were normalized.

The final stage is to integrate the adapted dictionary into the recognition or search framework. Many researchers make use of prior knowledge of pronunciation transformation to re-score the multiple hypotheses or lattices obtained in the original search process. In our work, a one-pass search mechanism was adopted: PDA information was utilized simultaneously with the language model and acoustic evaluation. This is illustrated with the following example.

Assume that speakers with a Shanghai accent probably utter "du2-bu4-yi1-shi2" (独步一时) as "du2-bu4-yi1-si2". The adapted dictionary could be as follows:

...		
shi2	shi2	0.83
shi2(2)	si2	0.17
...		
si2	si2	1.00
...		

Therefore, scores of the three partial paths yi1 -> shi2, yi1 -> shi2(2) and yi1 -> si2 could be computed respectively with formulae (1) (2) and (3).

$$\begin{aligned}&\text{Score}(\text{shi2}\mid\text{yi1})\\&\quad= w_{\text{LM}} * P_{\text{LM}}(\text{shi2}\mid\text{yi1}) + w_{\text{AM}}\\&\qquad * P_{\text{AM}}(\text{shi2}) + w_{\text{PDA}} * P_{\text{PDA}}(\text{shi2}\mid\text{shi2})\end{aligned} \tag{1}$$

$$\begin{aligned}&\text{Score}(\text{shi2(2)}\mid\text{yi1})\\&\quad= w_{\text{LM}} * P_{\text{LM}}(\text{shi2(2)}\mid\text{yi1}) + w_{\text{AM}} * P_{\text{AM}}(\text{shi2(2)})\\&\qquad + w_{\text{PDA}} * P_{\text{PDA}}(\text{shi2(2)}\mid\text{shi2})\\&\quad= w_{\text{LM}} * P_{\text{LM}}(\text{shi2}\mid\text{yi1}) + w_{\text{AM}} * P_{\text{AM}}(\text{si2})\\&\qquad + w_{\text{PDA}} * P_{\text{PDA}}(\text{si2}\mid\text{shi2})\end{aligned} \tag{2}$$

$$\begin{aligned}&\text{Score}(\text{si2}\mid\text{yi1})\\&\quad= w_{\text{LM}} * P_{\text{LM}}(\text{si2}\mid\text{yi1}) + w_{\text{AM}}\\&\qquad * P_{\text{AM}}(\text{si2}) + w_{\text{PDA}} * P_{\text{PDA}}(\text{si2}\mid\text{si2})\end{aligned} \tag{3}$$

where P_{LM}, P_{AM} and P_{PDA} stand for the logarithmic score of the language model (LM), acoustic model (AM) and pronunciation variation, respectively. w_{LM}, w_{AM} and w_{PDA} are the corresponding weight coefficients, which are usually determined empirically.

Obviously, the partial path yi1 -> shi2(2) has adopted the true pronunciation or acoustic model (as /si2/) while keeping the ought-to-be LM, e.g., bigram of (shi2 | yi1). At the same time, prior information about pronunciation transformation was incorporated. Theoretically, it should outscore the other two paths. As a result, the recognizer successfully recovers from a user's pronunciation mistake with PDA.

3.2. Experiments and Result

3.2.1. System Setup. Our baseline ASR system has been described in Section 2.1, but the training corpus

used is different. The acoustic model was trained on a database of 100,000 utterances collected from 250 male speakers from Beijing (BJ_Set), a subset of which was used to train model BJ as shown at Table 1. The baseline dictionary was an official published one that was consistent with the acoustic model. A tonal syllable trigram language model with perplexity of 98 on the test corpus was used in all experiments. Although the language model can compensate for some pronunciation discrepancies, experiments showed that PDA still significantly reduces recognition error of accented speech. Other data sets were as follows:

- Dictionary adaptation set (PDA_Set): 24 male speakers from the Shanghai region, 250 utterances per speaker, only 1/3 of the corpus (2000) utterances factually used.
- Testing set (Test_Set) 10 male speakers different from PDA_set, with Shanghai accent, 20 utterances per speaker, tagged with SH-M, as shown at Table 2;
- MLLR adaptation set (MLLR_Set): Same speaker set as testing set, 180 utterances that are different from Test_set per speaker.
- Accent-dependent set (SH_Set): 290 male speakers from Shanghai area, 250 utterances per speaker.
- Mixed accent set (MIX_Set): BJ_Set plus SH_Set.

3.2.2. Analysis. 2000 sentences from the PDA_Set were transcribed with the benchmark recognizer in term of standard sets and syllable loop grammar. Dynamic programming was applied to these results and some interesting linguistic phenomena were observed.

Front nasal and back nasal

Table 5 shows that final ING and IN are often exchangeable and ENG are often uttered as EN. However EN is seldom pronounced as ENG and not listed in the Table.

ZH (SH, CH) VS. Z (S, C)

Because of phonemic diversity, it is hard for Shanghai speakers to utter initial phonemes like /zh/, /ch/ and /sh/. As a result, syllables that include such phones are uttered into syllables initialized with /z/, /s/ and /c/, as shown in Table 6. Such results are in agreement with the philologists' studies.

3.2.3. Results. Recognition results with PDA, MLLR and the combination of the two are reported here. To illustrate the impact of different baseline systems on PDA and MLLR, the performance of the accent-dependent model (trained on SH_Set) and the accent-independent model (trained on Mix_Set) are also presented.

Table 5. Front/Back nasal mappings of accent speakers in term of standard pronunciations.

Canonical pron.	Observed pron.	Prob. (%)	Canonical pron.	Observed pron.	Prob. (%)
QIN	QING	47.37	QING	QIN	19.80
LIN	LING	41.67	LING	LIN	18.40
MIN	MING	36.00	MING	MIN	42.22
YIN	YING	35.23	YING	YIN	39.77
XIN	XING	33.73	XING	XIN	33.54
JIN	JING	32.86	JING	JIN	39.39
PIN	PING	32.20	PING	PIN	33.33
(IN)	**(ING)**	**37.0**	**(ING)**	**(IN)**	**32.4**
RENG	REN	55.56	SHENG	SHEN	40.49
GENG	GEN	51.72	CHENG	CHEN	25.49
ZHENG	ZHEN	46.27	NENG	NEN	24.56
MENG	MEN	40.74	**(ENG)**	**(EN)**	**40.7**

Table 6. ZH/SH/CH vs. Z/C/S mappings of accented speakers in term of standard pronunciations.

Canonical	Observed	Prob. (%)	Canonical	Observed	Prob.
ZHI	ZI	17.26	CHAO	CAO	37.50
SHI	SI	16.72	ZHAO	ZAO	29.79
CHI	CI	15.38	ZHONG	ZONG	24.71
ZHU	ZU	29.27	SHAN	SAN	19.23
SHU	SU	16.04	CHAN	CAN	17.95
CHU	CU	20.28	ZHANG	ZANG	17.82

3.2.3.1. PDA, MLLR, and the Combination of Both. Starting with many kinds of mapping pairs, we first removed pairs with few observations and low variation probabilities, and encoded the remaining ones into the dictionary. Table 7 shows the recognition result when we used 37 transformation pairs, consisting

Table 7. Performance of PDA (37 transformation pairs used for PDA).

Dictionary	Syllable error rate (%)
Baseline	23.18
+ PDA (w/o Prob.)	20.48 (+11.6%)
+PDA (with Prob.)	19.96 (+13.9%)

Table 8. Performance of MLLR and PDA/MLLR with different number of adaptation utterances.

No. of adp. utterances	0	10	20	30	45	90	180
MLLR	23.18	21.48	17.93	17.59	16.38	15.89	15.50
Rel. err. reduction	–	7.33	22.65	24.12	29.34	31.45	33.13
MLLR + PDA	19.96	21.12	17.50	16.59	15.77	15.22	14.83
Rel. err. reduction	13.89	8.89	24.50	28.43	31.97	34.34	36.02
Rel. err. reduction (on MLLR)	–	1.68	2.40	5.69	3.72	4.22	4.32

mainly of pairs shown in Tables 5 and 6. We tried two kinds of methods to deal with transformation pairs: without probability and with probability. The former factually assumes the same probability for both the canonical pronunciation and the alternative one. It is a method of simply introducing multiple pronunciations. The later method is more accurate to describe the pronunciation variations with real probabilities extracted from the accented corpora, as shown at Tables 5 and 6.

To evaluate the acoustic model adaptation performance, we also carried out a standard MLLR adaptation. All 187 phones were classified into 65 regression classes. Both diagonal matrix and bias offset are used in the MLLR transformation matrix. Adaptation sets sizes ranging from 10 to 180 utterances per testing speaker were used. Results are shown in the Table 8. The results show that when the number of adaptation utterances reaches 20, relative error reduction is more than 22%. Based on the assumption that PDA and MLLR can be complementary in pronunciation variation and acoustic characteristics, respectively, experiments combining MLLR and PDA were carried out. Compared with performance without adaptation, a 28.43% error reduction is achieved (30 adaptation utterances per speaker). Compared with MLLR alone, a further error reduction of 5.69% is obtained.

Table 9. Performance of PDA/MLLR based on different baselines (cross-accent (BJ_Set), accent-independent (Mix_Set) and accent-dependent (SH_Set)).

Different setup	Baseline (%)		
	BJ_Set	MIX_Set	SH_Set
Baseline	23.18	16.59	13.98
+PDA	19.96	15.56	13.76
+MLLR (30 Utts.)	17.59	14.40	13.49
+PDA+MLLR	16.59	14.31	13.52

3.2.3.2. Comparison of Different Models. Table 9 shows the results of PDA/MLLR based on three different baselines: cross-accent model, accent-dependent model and accent-independent model, trained on the BJ_set, SH_set and Mix_set, respectively, as shown at Table 9. The performance of PDA and/or MLLR increases with the distance between the baseline model available and the testing speakers. When the baseline did not include any accent information for the test speakers, PDA/MLLR has achieved the best results. When accent-independent model did include some training speakers with the same accent as the test speaker, PDA/MLLR still achieved positive, but not significant results. However, given an accent-dependent model, contributions of PDA/MLLR become marginal. In addition, the accent-dependent model still outperforms any other combinations of baseline models and adaptation methods. This motivated us to develop accent identification methods in those cases for which sufficient accent corpora exists.

4. Accent Identification for Accent-Dependent Model Selection

In some situations we can collect a large amount of data for each accent type, and thus accent-dependent (AD) models can be trained. As we observed in Sections 2 and 3, accent-dependent models always achieve the best performance. So, the remaining core problem for applying AD in recognition is the automatic identification of the accents of testing speakers given very little data.

Current accent identification research focuses on the foreign accent problem. That is, identifying non-native accents. Teixeira et al. (1996) proposed a HMM-based (Hidden Markov Model) system to identify English with six foreign accents: Danish, German, British, Spanish, Italian and Portuguese. A context-independent HMM was applied because the corpus

consisted of isolated words only, which is not always the case in applications. Hansen and Arslan (1995) also built a HMM to classify foreign accents of American English. They analyzed the impacts of prosodic features on classification performance and concluded that carefully selected prosodic features would improve classification accuracy. Instead of phoneme-based HMM, Fung and Liu (1999) used phoneme-class HMMs to differentiate Cantonese English from native English. Berkling et al. (1998) added English syllable structure knowledge to help recognize three accented speaker groups of Australian English.

Although foreign accent identification has been explored extensively, little has been done regarding domestic accents, to the best of our knowledge. Domestic accent identification is more challenging: (1) Some linguistic knowledge, such as syllable structure (used in Berkling et al., 1998), is of little use since people seldom make such mistakes in their mother language; and (2) Differences among domestic speakers are relatively smaller than these among foreign speakers. In our work, we are engaged in identifying different accent types spoken by people with the same mother tongue.

Most of current accent identification systems, as mentioned above, are built based on the HMM framework. Although HMM is effective in classifying accents, its training procedure is time-consuming. Also, using HMM to model every phoneme or phoneme-class is not efficient. Furthermore, HMM training is a supervised training that requires transcriptions. The transcriptions either are manually labeled, or obtained from a speech recognizer, in which case the word error rate degrades the identification performance.

In this section, we propose a GMM-based method for the identification of domestic speaker accents (Chen et al., 2001). GMM training is unsupervised: no transcriptions are needed. Four typical Mandarin accent types were explored: Beijing, Shanghai, Guangdong and Taiwan. We trained two GMMs for each accent: one for male, the other for female. Given test data, the speaker's gender and accent can be identified simultaneously, compared with the two-stage method discussed in Teixeira et al. (1996). The relationship between GMM parameter configurations and recognition accuracy was examined. We also investigated how many utterances per speaker were sufficient to recognize his/her accent reliably. We showed the correlations among accents, and provided some explanations. Finally, the efficiency of accent identification was also examined by applying it to speech recognition.

4.1. Multi-Accent Mandarin Corpus

The multi-accent Mandarin corpus, consisting of 1,440 speakers, is part of a corpora collected by Microsoft Research Asia. There are four accents: Beijing (BJ, including 3 channels: BJ, EW and FL), Shanghai (SH, including 2 channels: SH and JD), Guangdong (GD) and Taiwan (TW). All waveforms were recorded at a sampling rate of 16 KHz except for the TW ones, which were collected at 22 KHz and then downsampled. In the training corpus, there were 150 female and 150 male speakers of each accent with two utterances per speaker. In the test corpus, there were 30 female and 30 male speakers of each accent with 50 utterances per speaker. Most of the utterances lasted approximately 3–5 seconds each, forming about 16 hours' speech data for the entire corpus. There was no overlap of speakers and utterances the between training and test corpora.

4.2. Accent Identification System

Since gender and accent are important factors in speaker variability, the probability distributions of distorted features caused by different genders and accents are different. As a result, we used a set of GMMs to estimate the probability that the observed utterance came from a particular gender and accent.

In our work, M GMMs, $\{\Lambda_k\}_{k=1}^{M}$ are independently trained using the speech produced by a given gender and accent group. That is, model Λ_k is trained to maximize the log-likelihood function

$$\log \prod_{t=1}^{T} p(x(t) \mid \Lambda_k) = \sum_{t=1}^{T} \log p(x(t) \mid \Lambda_k), \quad k = 1, \ldots, M, \quad (4)$$

Where the speech feature is denoted by $x(t)$. T is the number of speech frames in the utterance and M is twice (two genders) the total number of accent types. The GMM parameters are estimated by the expectation maximization (EM) algorithm (Dempster et al., 1977). During identification, an utterance is fed to all the GMMs. The most likely gender and accent type is identified according to

$$\hat{k} = \arg\max_{k=1}^{M} \sum_{t=1}^{T} \log p(x(t) \mid \Lambda_k). \quad (5)$$

4.3. Experiments

As described in Section 4.1, there are eight subsets (accent plus gender) in the training corpora. In each subset, two utterances per speaker, altogether 300 utterances per subset, were used to train the GMMs. Because the 300 utterances are from 150 speakers with different ages, speaking rates and even recording channels, speaker variability caused by these factors was averaged. The test set consisted of 240 speakers from four accents with 50 utterances each. The features used were 39-order Mel-Frequency Cepstral Coefficients (MFCC), consisting of 12 cepstral coefficients, energy, and their first and second order differences. Cepstral mean subtraction was performed within each utterance to remove the effect of channels. Data preparation and training procedures were performed using the HTK 3.0 toolkit.

4.3.1. Number of Components in GMM. In this experiment, we examined the relationship between the number of components in GMMs and the identification accuracy. Since the eight subsets were labeled with gender and accent, our method identified speaker's gender and accent at the same time.

Table 10 and Fig. 4 show the gender and accent identification error rate, respectively, as a function of the number of components in GMMs. Table 10 shows that the gender identification error rate decreases significantly when components increase from 8 to 32. However, only a small improvement is gained by using 64 components, as compared with 32. It can be concluded that a GMM with 32 components is capable of modeling gender variability of speech signals effectively.

Figure 4 shows a similar trend. It is clear that the number of components in GMMs greatly affects the accent identification performance. In contrast with the gender experiment, for accent, GMMs with 64 components still gain some improvement over 32-component GMMs (Error rate decreases from 19.1% to 16.8%).

Table 10. Gender identification vs. number of GMM components (four utterances used per speaker, relative error reduction is calculated when regarding GMM with eight components as the baseline).

No. of components	8	16	32	64
Error rate (%)	8.5	4.5	3.4	3.0
Rel. err. reduction (%)	–	47.1	60.0	64.7

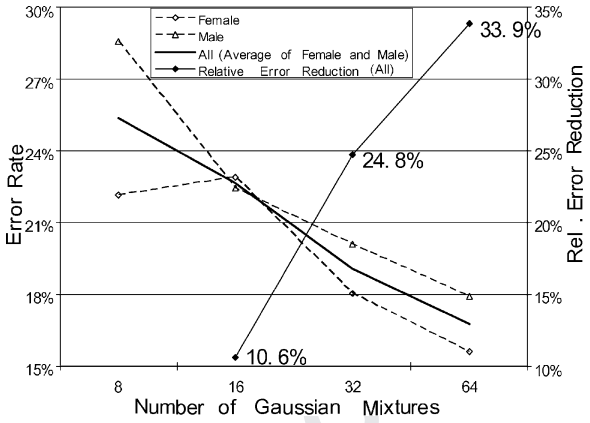

Figure 4. Accent identification error rate vs. different number of components of GMM. The right Y axis is the relative error reduction to eight components, when regarding GMM with eight components as the baseline. "All" means error rate averaged between females and males.

It is probably due to there are larger variances among accents types than that of gender.

Considering the training effort and the reliability of estimations, GMMs with 32 components are a good tradeoff and are used in the following experiments.

4.3.2. Number of Utterances per Speaker. In this experiment, we were concerned with the robustness of the method: how many utterances are sufficient to classify accent types reliably. We randomly selected N ($N \leq 50$) utterances for each test speaker and averaged their log-likelihood in each GMM. The test speaker was classified into the subset with the largest averaged log-likelihood. The random selection was repeated ten times to guarantee the achievement of reliable results.

Table 11 and Fig. 5 show the gender and accent identification error rates, respectively, varying the number of utterances. When averaging the log-likelihood of all 50 utterances of a speaker, there was no need to perform random selection.

Table 11 shows that gender identification is more reliable when more utterances are used. When the num-

Table 11. Gender identification vs. number of testing utterances (32 components/GMM used, relative error reduction is calculated when regarding one utterance as the baseline).

No. of utterances	1	2	3	4	5	10	20	50
Error rate (%)	3.4	2.8	2.5	2.2	2.3	1.9	2.0	1.2
Rel. error reduction (%)	–	18	26	35	32	44	41	65

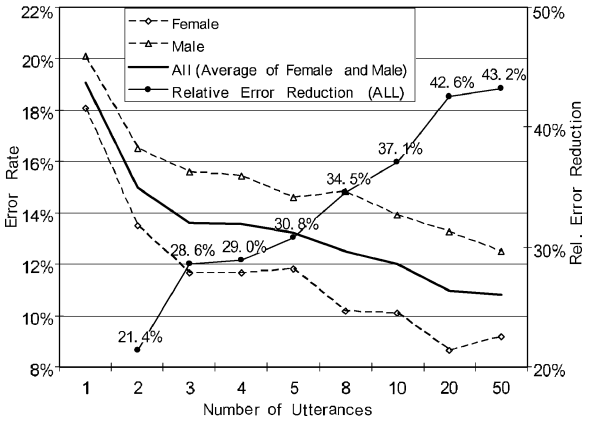

Figure 5. Accent identification error rate vs. different number of testing utterances per speaker. The right Y axis is the relative error reduction regarding one utterance as the baseline. "All" means error rate averaged between females and males.

ber of utterances increased from one to four, the gender identification errors were greatly reduced (35%). Further improvement was observed when using more than ten utterances, but it is not practical to collect so much data in many applications. As a tradeoff, 3–5 utterances are good enough in most situations.

It is clear from Fig. 5 that increasing the number of utterances improved the identification of accents. This is consistent with our intuition that more utterances from a speaker help in identifying his/her accent. Considering the tradeoff between accuracy and costs, using 3–5 utterances is a good choice, with an error rate of 13.2%–13.6%.

4.3.3. Discussions on Inter-Accent Results. To investigate the internal relationships among four accent types, we used the experiment based on 32 components and four utterances per testing speaker as a case, as illustrated in Table 12. Some findings are discussed as follows:

Table 12. Accents identification confusion matrices (32 components/GMM and four utterances per testing speaker).

	Testing utterances from			
Recognized as	BJ	SH	GD	TW
BJ	**0.775**	0.081	0.037	0.001
SH	0.120	**0.812**	0.076	0.014
GD	0.105	0.105	**0.886**	0.000
TW	0.000	0.002	0.001	**0.985**

- Compared with Beijing and Taiwan, Shanghai and Guangdong are likely to be misrecognized mutually, except themselves. In fact, Shanghai and Guangdong both belong to the southern language tree in phonology and share some common characteristics. For example, they do not differentiate front nasal and back nasal.
- The excellent performance for Taiwan speakers may have occurred for two reasons. First, Taiwan civilians may present some specialty on pronunciations from mainland due to regions distance. Second, limited by the recording condition, there is a certain portion of noise in the waveforms of the Taiwan corpora (both training and test) that makes them more distinct from the other accent types.
- The reason for the relatively low accuracy of Beijing possibly may be due to larger channel variations existing in the corpora. There are three channels in the Beijing corpus while there are two in the Shanghai corpus and one for Guangdong and Taiwan.
- Channel effects may constitute a considerable factor in GMM-based accent identification systems. For Beijing, Shanghai and Guangdong, accuracy decreased with an increasing number of channels. Further work is needed to weaken this effect.

4.3.4. Accent Dependent Model. In this subsection, we verify the efficiency of applying an accent-dependent model in speech recognition. Here, the baseline ASR system is the same as that used in the cross-accent experiments (Section 2.1). Considering three accent types (Beijing (BJ), Shanghai (SH) and Guangdong (GD)), there are six gender/accent-dependent acoustic models. Test sets were the same as that in Section 2.1, except that ten utterances per speaker were used for testing while the remaining ten utterances were used for MLLR adaptation (four utterances for selecting the right model). As shown in Table 13, AD models selected by automatic accent

Table 13. Performance of automatic accent identification in terms of speech recognition accuracy (Number shown in this table are measured with character error rate (%)).

	BJ	SH	GD
Baseline (X6)	9.07	17.86	30.42
AD (manually labeled accent)	9.16	16.62	25.72
AD (identified accent by GMM)	9.55	17.37	26.28
MLLR (1 class, diagonal + bias)	9.47	17.97	30.70

identification achieved comparable results to those that were manually labeled, especially for GD, which has only 1/6 the data of X6. The remaining gap is due mainly to incorrectly selected AD models rather than gender-dependent model. Relative to the baseline, no improvements were observed with MLLR adaptation of ten utterances per speaker.

5. Conclusion

It is widely known that speaker variability affects speech recognition performance greatly. It is also intuitive that accent is one of the main factors that causes variability and should impact the recognition. But what are the real effects and how should we deal with the problem in a real recognizer? In this paper, we first confirm this issue both quantitatively and qualitatively. Specifically, we carried out extensive experiments to evaluate the effect of accent on speech recognition, based on a state-of-the-art recognizer, and showed a 40–50% error increase for cross-accent speech recognition. Then, a high-level analysis based on PCA/ICA confirmed qualitatively that accent is another dominant factor, in addition to gender, in speaker variability.

Based on the above investigations, we explore this problem in two directions:

- Pronunciation adaptation. A pronunciation dictionary adaptation (PDA) method was proposed to model the pronunciation variation between speakers with standard pronunciation and the accented speakers. In addition to pronunciation level adjustments, we also applied acoustic level adaptation techniques, such as MLLR and an integration of both PDA and MLLR. PDA can deal with most dominant variation among accents group at the phonology level, while general speaker adaptation can trace the detailed changes of specific speakers such as speaking speed and style, at the acoustic level. Result shows that they are complementary.
- Building accent-dependent models and automatic accent identification. In cases where there are enough training data for each accent, more specific models can be trained with less speaker variability. In the paper, we proposed a GMM-based automatic detection method for regional accents. Compared with HMM-based identification methods, there is no need to know the transcription in advance because the training is text-independent. Also, the model size of a GMM is much more compact than that of a HMM. Therefore, there is much less training effort for a GMM, and its decoding procedure also is more efficient. The efficiency of accent identification in selecting accent-dependent models in recognition is supported by our experiments.

These two methods can be adopted in different cases according to available corpora. Given an amount of accented utterances insufficient to train an accent-specific model, we can extract the main pronunciation variations between accent groups and standard speakers through PDA. Without any changes at the acoustic and language model levels, pronunciation dictionaries can be adapted to deal with accented speakers. When a large amount of corpora for different accents can be obtained, accent-specific models can be trained and applied in speech recognition through a GMM-based automatic accent identification strategy.

The second method can be extended to more general cases, in addition to accent and gender. Given more detailed-labeled data such as speaking rate, we can train a speed-specific model in addition to the accent/gender-specific model to fit accurately to the speaker. The automatic identification strategy also can be used to cluster the huge amount of untagged data into more subsets and form clustered models. Then the right model can be selected with the same strategy. This strategy is applicable especially in client-server based applications where there are fewer limitations of space and computation. In this case, incrementally collected data can be classified and formed into more and more specific clustered models. The final target speaker model can be selected or adaptively combined (Gales, 2000) from multiple models.

Currently we are trying some new speaker representation methods for more efficient analysis of speaker variability (Chen et al., 2002). We are also introducing the GMM-based automatic identification strategy into unsupervised clustering training. Furthermore, a general speaker adaptation method, namely speaker selection training (Huang et al., 2002), which includes accent adaptation, is under development for fast adaptation

Acknowledgment

The authors thank the three anonymous reviewers for their critical reading of the manuscript and for their valuable comments and suggestions for improving the quality of the paper.

Notes

1. Accent, as addressed in this paper, is determined by the phonetic habits of the speaker's dialect carried over to his or her use of the mother tongue. In particular, it refers to speaking Mandarin with different regional accents caused by dialects, such as Shanghai and Guangdong.
2. PPC is a measure similar to word perplexity (PPW) except that it is based on character level. Usually PPW $=$ PPC$_{\wedge}$n, where n is the average word length in terms of characters of test corpora with a given lexicon.
3. Compared with the rest of the phone classes, these single vowels can reflect the speakers' characteristics efficiently. In addition, they are widely used and therefore can be estimated reliably. As a result, regressions classes corresponded to them are chosen as "supportive" representations of speakers.
4. There are eight major dialectal regions in China in addition to Mandarin (Northern China).They are called Wu, Xiang, Hui, Gan, Min, Jin, Hakka and Yue. The BJ, SH, GD and TW we discuss in this paper are speakers mainly from Mandarin, Wu, Yue and Min dialectal regions, respectively and they can and are required to speak Mandarin in the paper.
5. Syllable in Mandarin is a complete unit to describe a pronunciation of a Chinese character. It usually has five different tones and consists of two parts, called initial and final. E.g. /shi4/, /shi/ is the base syllable, 4 is the tone part and /sh/ is the initial and /i/ is the final.

References

Berkling, K., Zissman, M., Vonwiller, J., and Cleirigh, C. (1998). Improving accent identification through knowledge of English syllable structure. *Proc. International Conference on Spoken Language Processing*, vol. 2, pp. 89–92.

Chang, E., Zhou, J., Huang, C., Di, S., and Lee, K.F. (2000). Large vocabulary mandarin speech recognition with different approaches in modeling tones. *Proc. International Conference on Spoken Language Processing*, vol. 2, pp. 983–986.

Chen, T., Huang, C., Chang, E., and Wang, J. (2001). Automatic accent identification using Gaussian mixture models. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Italy.

Chen, T., Huang, C., Chang, E., and Wang, J. (2002). On the use of Gaussian mixture model for speaker variability analysis. *Proc. International Conference on Spoken Language Processing*, vol. 2, pp. 1249–1252.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*:1–38.

Fung, P. and Liu, W.K. (1999). Fast accent identification and accented speech recognition. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 221–224.

Gales, M.J.F. (2000). Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, *8*:417–428.

Hansen, J.H.L. and Arslan, L.M. (1995). Foreign accent classification using source generator based prosodic features. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 836–839.

Hotellings, H. (1933). Analysis of a complex of statistical variables into principle components. *J. Educ. Psychol.*, *24*:417–441, 498–520.

Hu, Z.H. (1999). Understanding and adapting to speaker variability using correlation-based principal component analysis. PhD Dissertation, Oregon Graduate Institute of Science and Technology.

Huang, C., Chang, E., Zhou, J.L., and Lee, K.F. (2000). Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. *Proc. International Conference on Spoken Language Processing*, vol. 3, pp. 818–821.

Huang, C., Chen, T., Li, S., Chang, E., and Zhou, J.L. (2001). Analysis of speaker variability. *Proc. European Conference on Speech Communication and Technology*. Denmark, vol. 2, pp. 1377–1380.

Huang, C., Chen, T., and Chang, E. (2002) Speaker selection training for large vocabulary continuous speech recognition, *Proc. International Conference on Acoustics, Speech, and Signal Processing*. Florida, USA. vol. 1, pp. 609–612.

Humphries, J.J. and Woodland, P.C. (1998). The use of accent-specific pronunciation dictionaries in acoustic model training. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 317–320.

Hyvarinen, A. and Oja, E. (2000). Independent component analysis: algorithms and application. *Neural Networks*, *13*:411–430.

Lee, C.-H., Lin C.-H., and Juang, B.-H. (1991). A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Transactions on Signal Processing*, *39*:806–814.

Leggetter, C.J. and Woodland, P.C. (1995). Maximum likely-hood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, *9*:171–185.

Liu, M.K., Xu, B., Huang, T.Y., Deng, Y.G., and Li, C.R. (2000). Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1025–1028.

Malayath, N., Hermansky, H., and Kain, A. (1997). Towards decomposing the sources of variability in speech. *Proc. European Conference on Speech Communication and Technology*, vol. 1, pp. 497–500.

Riley, M.D. and Ljolje, A. (1996). Automatic generation of detailed pronunciation lexicon. *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer Academic Press, ch. 12, pp. 285–302.

Riley, M.D., 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-labelled phonetic corpora. *Speech Communication*, *29*:209–224.

Strik, H. and Cucchiarini, C. (1998) Modeling pronunciation variation for ASR: Overview and comparison of methods. *Proc. ETRW Workshop on Modeling Pronunciation Variation for ASR*, Kerkrade, pp. 137–144 .

Teixeira, C., Trancoso, I., and Serralheiro, A. (1996). Accent identification. *Proc. International Conference on Spoken Language Processing*, vol. 3, pp. 1784–1787.