Accent Issues in Large Vocabulary Continuous Speech Recognition (LVCSR)

Chao Huang
Eric Chang
Tao Chen

August 8, 2001

Technical Report
MSR-TR-2001-69

Microsoft Research China
Microsoft Corporation
5F, Beijing Sigma Center,
No. 49, Zhichun Road, Haidian District
Beijing 100080, P.R. China

Abstract

Speech recognition has achieved great improvements recently. However, robustness is still one of the big problems, e.g. performance of recognition fluctuates sharply depending on the speaker, especially when the speaker has strong accent that is not covered in the training corpus. In this report, we first introduce our result on cross accent experiments and show a 30% error rate increase when accent independent models are used instead of accent dependent ones. Then we organize the report into three parts to cover the problem. In the first part, we do an investigation of speaker variability and manage to seek out the relationship between the well-known parameter representation and the physical characteristics of speaker, especially accent and confirm once more that accent is one of the main factors causing speaker variability. Then we provide our solutions for accent variability from two aspects. One is adaptation method, including pronunciation dictionary adaptation and acoustic model adaptation, which integrate the dominant changes among accent speaker groups and the detailed style for specific speaker in each group. The other is to build accent specific models as we do in cross accent experiments. The key point inside this method is to provide an automatic mechanism to choose the accent dependent model, which is explored in the fourth part of the report. We propose a fast and efficient GMM based accent identification method. The respective descriptions of three parts are outlined as follows.

Analysis and modeling of speaker variability, such as gender, accent, age, speaking rate, and phone realizations, are important issues in speech recognition. It is known that existing feature representations describing speaker variations are high dimensional. In the third part of this report, we introduce two powerful multivariate statistical analysis methods, namely, principal component analysis (PCA) and independent component analysis (ICA), as tools to analyze such variability and extract low dimensional feature representation. Our findings are the following: (1) the first two principal components correspond to gender and accent, respectively. (2) It is shown that ICA based features yield better classification performance than PCA ones. Using 2-dimensional ICA representation, we achieve 6.1% and 13.3% error rate in gender and accent classification, respectively, for 980 speakers.

In the fourth part, a method of accent modeling through Pronunciation Dictionary Adaptation (PDA) is presented. We derive the pronunciation variation between canonical speaker groups and accent groups and add an encoding of the differences to a canonical dictionary to create a new, adapted dictionary that reflects the accent characteristics. The pronunciation variation information is then integrated with acoustic and language models into a one-pass search framework. It is assumed that acoustic deviation and pronunciation variation are independent but complementary phenomena that cause poor performance among accented speakers. Therefore, MLLR, an efficient model adaptation technique, is also presented both alone and in combination with PDA. It is shown that when PDA, MLLR and the combination of them are used, error rate reductions of 13.9%, 24.1% and 28.4% respectively, are achieved.

It is well known that speaker variability caused by accent is an important factor in speech recognition. Some major accents in China are so different as to make this problem very severe. In part 5, we propose a Gaussian mixture model (GMM) based Mandarin accent identification method. In this method, a number of GMMs are trained to identify the most likely accent given test utterances. The identified accent type can be used to select an accent-dependent model for speech recognition. A multi-accent Mandarin corpus was developed for the task, including 4 typical accents in China with 1,440 speakers (1,200 for training, 240 for testing). We explore experimentally the effect of the number of components in GMM on identification performance. We also investigate how many utterances per speaker are sufficient to reliably recognize his/her accent. Finally, we show the correlations among accents and provide some discussions.

Keywords: Accent modeling, automatic accent identification, pronunciation dictionary adaptation (PDA), speaker variability, principal component analysis (PCA), independent component analysis (ICA).

Table of Contents

List of Figures

List of Tables

1. Introduction

It is well known that state-of-the-art speech recognition (SR) systems, even in the domain of large vocabulary continuous speech recognition, have achieved great improvements in last decades. There are several commercial systems on the shelves like ViaVoice of IBM, SAPI of Microsoft and FreeSpeech of Phillips.

Speaker variability greatly impacts the performance of SR. Among the variability, gender and accent are two most important factors that cause the variance among speakers [12]. The former has been lift up by the gender dependent models. However, there is comparatively little research on the topic of accented speech recognition, especially when the speakers come from the same mother tongue, but with accents because of the different dialects.

In this report, firstly, we will explore the impact of accented speech on recognition performance. According to our experiments, there is 30% relative error increase when speech is mixed with accent. We investigate the problem from two different views: accent adaptation through pronunciation dictionary adaptation (PDA) that built for specific accents and accent specific modeling training on the acoustic levels. We will briefly introduce the two strategies after some data-driven analysis of speaker variability.

In the second part of the report, we make a detailed investigation about speaker variability, specifically on gender and accent. The motivation is to establish the relationship between the dominant feature representations of current speech recognition systems and the physical characteristics of speakers, such as accent and gender. It is shown that accent is the second greatest factor among speaker variability [12]. This motivates us to look for strategies to solve this problem.

In the first strategy to deal with accent problem, PDA [18] tries to seek the pronunciation variations among speakers coming from different accents and model such difference on the dictionary level. In practice, we often adopt the well-known pronunciations as the baseline system, and then extract the pronunciation changes through speaker-independent system or phonology rules. Finally we encode such changes into reference dictionary and obtain an accent specific dictionary. The variations may be mapping pairs of phone, phoneme, or syllable including substitution, insertion and deletion. These mapping rules can be learned automatically through some enrollments of accented speech recognized by baseline recognition system or summarization of phonologies. In addition, it can be context dependent or independent.

The second strategy is to build the accent-specific model, which is easy to be understood. Sufficient corpus is necessary for each accent set. Just like the gender-dependent model, accent dependent models can greatly reduce the variance of the each separated set and thus improve the performance, which will be confirmed in the following sections. Although it is probably not efficient to provide multiple model sets in the desktop-based application, it is practical when the application is built on the client-server structure.

However, the core problem of such strategy is to select the proper model for each target speaker. In other words, a method to identify the incoming speaker's characteristics such as gender and accent automatically in order to choose the corresponding model is important and very meaningful. We proposed a Gaussian Mixture Model (GMM) based accent (including gender) identification method. In our work, M GMMs, $\{\Lambda_k\}_{k=1}^{M}$, are independently trained using the speech produced by the corresponding gender and accent group. That is, model Λ_k is trained to maximize the log-likelihood function

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

Where the speech feature is denoted by $x(t)$. T is the number of speech frames in the utterance and M is twice (male and female) the total number of accent types. The GMM parameters are estimated by the expectation maximization (EM) algorithm [17]. During identification, an utterance is fed to all the GMMs. The most likely gender and accent type is identified according to

$$\hat{k}=\arg\max_{k=1}^{M}\sum_{t=1}^{T}\log p(x(t)\mid\Lambda_k). \qquad (2)$$

In this report, some main accents of Mandarin, including Beijing, Shanghai, Guangdong and Taiwan are considered.

2. Cross Accent Experiments

In order to investigate the impact of accent on the state of the art speech recognition system, we have carried lots of experiments based on Microsoft Chinese speech engine, which has been successfully delivered into Office XP and SAPI. In addition to many kinds of mature technologies such as Cepstrum Mean Normalization, decision tree based state tying, context dependent modeling (triphone) and trigram language modeling, which are all been testified to be important and adopted in the system, tone related information, which are very helpful to be distinguished for Asian tonal language, have also been integrated into out baseline system through including pitch and delta pitch into feature streams and detailed tone modeling. In one word, all improvements and results shown here are achieved based on a solid and powerful baseline system.

The details about experiment and results are listed as follows:

Experiments setup

- Training corpus and model configurations

Table 2.1: Summary of training corpora for cross accent experiments, Here BJ, SH and GD means Beijing, Shanghai and Guangdong accent respectively.

Model Tag	Training corpus configurations	Accent specific model
EW	500BJ	BJ
BEF	~1500BJ	BJ
JS	~1000SH	SH
GD	~500GD	GD
BES	~1000BJ+ ~500SH	Mixed (BJ+SH)
X5	~1500BJ+ ~1000SH	Mixed (BJ+SH))
X6	~1500BJ+ ~1000SH+ ~500GD	Mixed (BJ+SH+GD)

- Test corpus

Table 2.2: Summary of test corpora for cross accent experiments, PPc show here is character perplexity of test corpora according to the LM of 54K.Dic and BG=TG=300,000.

Test Sets	Accent	Speakers	Utterances	Characters	PPc
m-msr	Beijing	25	500	9570	33.7
f-msr	Beijing	25	500	9423	
m-863b	Beijing	30	300	3797	41.0
f-863b	Beijing	30	300	3713	
m-sh	Shanghai	10	200	3243	59.1
f-sh	Shanghai	10	200	3287	
m-gd	Guangdong	10	200	3233	55-60
f-gd	Guangdong	10	200	3294	
m_it	Mixed (mainly Beijing)	50	1,000	13,804	
f-it	Mixed (mainly Beijing)	50	1,000	13,791	

- Experiments Result

Table 2.3: Character error rate for cross accent experiments.

Model	Different accent test sets				
	MSR	**863**	**SH**	**GD**	**IT**
EW(500BJ)	9.49	11.89	22.67	33.77	19.96
BEF(1500BJ)	8.81	**10.80**	21.85	31.92	19.58
JS(1000SH)	10.61	13.89	**15.64**	28.44	22.76
GD(500GD)	12.94	13.96	18.71	**21.75**	28.28
BES(1000BJ+500SH)	**8.56**	10.85	18.14	30.19	**19.42**
X5(1500BJ+1000SH)	8.87	10.95	16.80	29.24	19.78
X6(1500BJ+1000SH+500GD)	9.02		17.59	27.95	

It is easily concluded from Table 2.3 that accent is a big problem that impacts the state of the art speech recognition systems. Compared with accent specific model, cross accent model may increase error rate by 40-50%.

3. Investigation of Speaker Variability

3.1 Introduction

Speaker variability, such as gender, accent, age, speaking rate, and phones realizations, is one of the main difficulties in speech signals. How they correlate each other and what the key factors are in speech realization are real concerns in speech research. As we know, performance of speaker-independent (SI) recognition systems is generally 2-3 times worse than that of speaker-dependent ones. As an alternative, different adaptation techniques, such as MAP and MLLR, have been used. The basic idea is to adjust the SI model and make it reflect intrinsic characteristics about specific speakers by re-training the system using appropriate corpora. Another method to deal with the speaker variability problem is to build multiple models of smaller variances, such as gender dependent model and accent dependent model, and then use a proper model selection scheme for the adaptation. SI system and speaker adaptation can be facilitated if the principal variances can be modeled and corresponding compensations can be made.

Another difficulty in speech recognition is the complexity of speech models. There can be a huge number of free parameters associated with a set of models. In other words, a representation of a speaker has to be high-dimensional when different phones are taken into account. How to analyze such data is a challenge.

Fortunately, several powerful tools, such as principal component analysis (PCA) [2] and more recently independent component analysis (ICA) [1], are available for high dimension multivariate statistical analysis. They have been applied widely and successfully in many research fields such as pattern recognition, learning and image analysis. Recent years have seen some applications in speech analysis [4][5][6].

PCA decorrelates second order moments corresponding to low frequency property and extracts orthogonal principal components of variations. ICA is a linear, not necessarily orthogonal, transform which 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 dependencies. It extracts independent components even if their magnitudes are small whereas PCA extracts components having largest magnitudes. ICA representation seems to capture the essential structure of the data in many applications including feature extraction and signal separation

In this section, we present a subspace analysis method for the analysis of speaker variability and for the extraction of low-dimensional speech features. The transformation matrix obtained by using maximum likelihood linear regression (MLLR) is adopted as the original representation of the speaker characteristics. Generally each speaker is a

super-vector which includes different regression classes (65 classes at most), with each class being a vector. Important components in a low-dimensional space are extracted as the result of PCA or ICA. We find that the first two principal components clearly present the characteristics about the gender and accent, respectively. That the second component corresponds to accent has never been reported before, while it has been shown that the first component corresponds to gender [5][6]. Furthermore, using ICA features can improve classification performance than using PCA ones. Using the ICA representation and a simple threshold method, we achieve gender classification accuracy of 93.9% and accent accuracy of 86.7% for a data set of 980 speakers.

3.2 Speaker Variance Investigations

3.2.1 Related Work

PCA and ICA have been widely used in image processing, especially in face recognition, identification and tracing. However, their application in speech field is comparatively rare. Like linear discriminant analysis (LDA), most speech researchers use PCA to extract or select the acoustic features. [4]. Kuhn et al. applied PCA at the level of speaker representation and proposed eigenvoices in analog to eigenfaces and further apply it in the rapid speaker adaptation [5]. Hu applied PCA to vowel classification [6].

All above work are based on representing speakers with concatenate the mean feature vector of vowels [6] or put one line of all the means from the Gaussian model that specifically trained for a certain speaker [5]. We have adopted the speaker adaptation model, specifically; we use the transformation matrix and offset that are adapted from the speaker independent model to represent the speaker. Here, maximum likelihood linear regression (MLLR) [11] was used in our experiments.

In addition, all above work only use PCA to pursue the projection of speaker in low dimension space in order to classify the vowels or construct the speaker space efficiently. As we know, PCA uses only second-order statistics and emphasize the dimension reduction, while ICA depends on the high-order statistics other than second order. PCA is mainly aim to the Gaussian data and ICA aiming to the Non-Gaussian data. Therefore, based on PCA, we introduce ICA to analysis the variability of speaker further because we have no clear sense on the statistical characteristics of speaker variability initially.

3.2.2 Speaker Representation

MLLR Matrices vs. Gaussian Models

As mentioned in Section 3.2.1, we have used the MLLR transformation matrix (including offset) to represent all the characteristics of a speaker, instead of using the means of the Gaussian models. The main advantage is such a representation provides a flexible means to control the model parameters according to the available adaptation corpora. The baseline system and setups can be found in [3]. To reflect the speaker in detail, we have tried to use multiple regression classes, at most 65 according to the phonetic structures of Mandarin.

Supporting Regression Classes Selection

We have used two different strategies to remove undesirable effects brought about by different phones. The first is to use the matrices of all regression classes. However, this increases the number of parameters that have to be estimated and hence increases the burden on the requirements on the adaptation corpora. In the second strategy, we choose empirically several supporting regression classes among all. This leads to significant decrease in the number of parameters to be estimated; and when the regression classes are chosen properly, there is little sacrifice in accuracy; as will be shown in Tables 3.4 and 3.5 in Section 3.3. The benefit is mainly due to that a proper set of support regression classes are good representatives of speakers in the sense that they provide good discriminative feature for the classification between speakers. Furthermore, fewer classes mean lower degree of freedom and increase in the reliability of parameters.

Diagonal Matrix vs. Offsets

Both diagonal matrix and offset are considered when making the MLLR adaptation. We have experimented with three combinations to represent speakers in this level: only diagonal matrix (with tag d), only offset (with tag b) and both of them (with tag bd). The only offset item of MLLR transformation matrix achieved much better result in gender classification, as will be shown in Table 3.3.

Acoustic Feature Pruning

The state of art speech recognition systems often apply multiple order dynamic features, such as first-order difference and second-order one, in addition to the cepstrum and energy. However, the main purpose of doing so is to build the speaker independent system. Usually, the less speaker-dependent information is involved in the training process, the better the final result will be. In contrast to such a feature selection strategy, we choose to extract the speaker-dependent features and use them to effectively represent speaker variability. We have applied several pruning strategies in the acoustic features level. We have also integrated pitch related features into our feature streams. Therefore, there are the six feature pruning methods as summarized in Table 3.1.

Table 3.1: Different feature pruning methods (number in each cell mean the finally kept dimensions used to represent the speaker).

Dynamic features	0-order (static)	1 order	2 order
w/o pitch	13	26	33
w/ pitch	14	28	36

3.3 Experiments

3.3.1 Data Corpora and SI Model

The whole corpora contain 980 speakers, 200 utterances per speaker. They are from two accent areas in China, Beijing (EW) and Shanghai (SH). The gender and accent distributions are summarized in Table 3.2.

Table 3.2: Distribution of speakers in corpora**.**

	Beijing	Shanghai
Female	250 (EW-f)	190 (SH-f)
Male	250 (EW-m)	290 (SH-m)

The speaker-independent model we used to extract the MLLR matrix is trained according to all corpora from EW. It is also gender–independent, unlike the baseline system.

3.3.2 Efficient Speaker Representation

Figure 3.1 show the component contribution and cumulative contribution of top *N* principal components on variances, where *N*=1, 2…156. The PCA algorithm used in these and the following experiments is based on the covariance matrix. The dynamic range for each dimension has been normalized for each sample. This way, covariance matrix becomes the same as the correlation matrix.

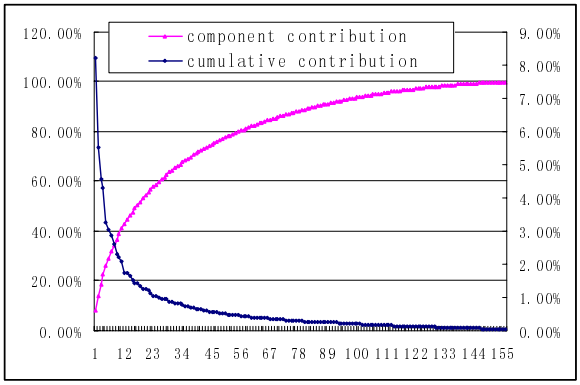

Figure 3.1: Single and cumulative variance contribution of top N components in PCA (horizontal axis means the eigenvalues order, left vertical axis mean cumulative contribution and right vertical axis mean single contribution for each eigenvalue).

To find the efficient and typical representation about speaker characteristics, we have applied strategies at several levels from supporting regression classes to acoustic features. Table 3.3 shows the gender classification results based on EW and SH corpora for various methods. Tags of -b,-d and –bd in the first column are according to the definition in section 2.2.3. Here the number of supporting regression classes is 6. From Table 3.3, we can conclude that the offset item in the MLLR matrix gives the best result.

Furthermore, among all the acoustic feature combinations, the combination of the static features, first order of cepstrum and energy gives the best result for both EW and SH sets. It can be explained that these dimensions carry the most of the speaker specific information. However, it is very interesting to note that the addition of the pitch related dimensions leads to a slight decrease in the accuracy. It contradicts to the common conclusion that the pitch itself is the most significant feature of gender. This may be due

to the following two reasons: First, pitch used here is in the model transformation level instead of the feature level. Secondly, multiple–order cepstrum feature dimensions have already included speaker gender information.

Table 3.3: Gender classifications errors based on different speaker representation methods (The result is according to the projection of PCA, the total number for EW and SH are 500 and 480 respectively).

Dims	13	26	33	14	28	36
SH-b	22	14	24	22	20	30
SH-d	58	78	80	62	82	86
SH-bd	34	42	46	38	40	46
EW-b	52	38	66	52	56	78
EW-d	76	124	100	108	140	118
EW-bd	48	92	128	88	82	122

To evaluate the proposed strategy for the selection of supporting regression classes, we made the following experiments. There are a total of 65 classes. Here only the offset of MLLR transformation matrix and the 26 dimensions in feature stream are used according to the results demonstrated in Table 3.3. The selections of different regression classes are defined in Table 3.4, and the corresponding gender classification results are shown in Table 3.5.

Obviously, the combination of the 6 regression classes is a proper choice to balance the classification accuracy and the number of model parameters. Therefore, in the following experiments where the physical meaning of the top projections is investigated, we optimize the input speaker representation with the following setups:

- Supporting regression classes: 6 single vowels (/a/, /i/, /o/, /e/, /u/, /v/)
- Offset item in MLLR transformation matrix;
- 26 dimensions in acoustic feature level

As a result, a speaker is typically represented with a supervector of 6*1*26=156 dimension.

Table 3.4: Different supporting regression classes selection.

# of regression classes	Descriptions
65	All classes
38	All classes of finals
27	All classes of initials
6	/a/, /i/, /o/, /e/, /u/, /v/
3	/a/, /i/, /u/
2	/a/, /i/
1	/a/

Table 3.5: Gender classifications errors of EW based on different supporting regression classes (The relative size of feature vector length is indicated as Parameters).

Number of Regression Classes	65	38	27	6	3	2	1
Errors	32	36	56	38	98	150	140
Parameter	--	0.58	0.42	0.09	0.046	0.03	0.015

3.3.3 Speaker Space and Physical Interpretations

The experiments here are performed with the mixed corpora sets of EW and SH. In this case, the PCA is performed with 980 samples of 156 dimensions each. Then, all speakers are projected into the top 6 components. A matrix of 980* 6 is obtained and is used as the input to ICA (The ICA is implemented according to the algorithm of FastICA proposed by Hyvarinen [1]). Figure 3.2 and Figure 3.3 show the projections of all the data onto the first two independent components. In the horizontal direction is the speaker index for the two sets. The alignment is: EW-f (1-250), SH-f (251-440), EW-m (441-690) and SH-m (691-980).

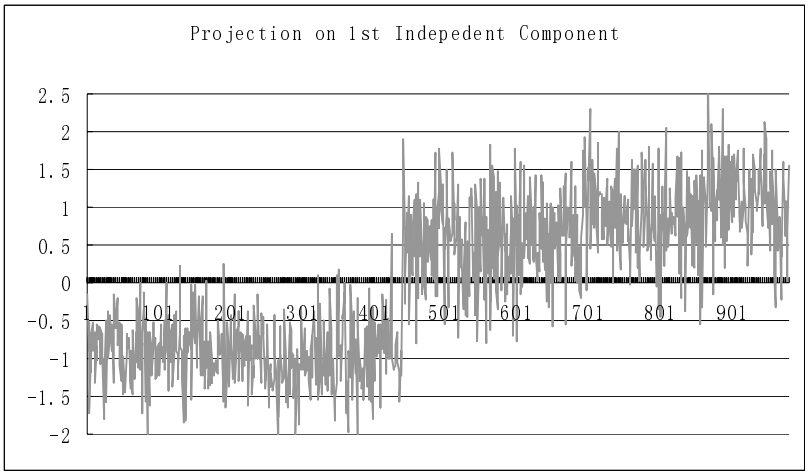

Figure 3.2: Projection of all speakers on the first independent component (The first block corresponds to the speaker sets of EW-f and SH-f, and the second block corresponds to the EW-m and SH-m).

From Figure 3.2, we can make a clear conclusion that the independent component corresponds to the gender characteristics of speaker. Projections on this component almost separate all speakers into two categories: male and female.

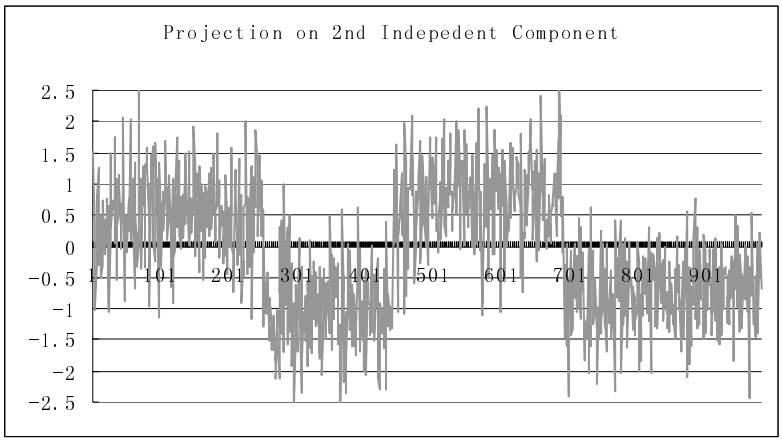

Figure 3.3: Projections of all speakers on the second independent component. (The four blocks correspond to the speaker sets of EW-f, SH-f, EW-m, SH-m from left to right).

According to Figure 3.3, four subsets occupy four blocks. The first and the third one together correspond the accent set EW (with Beijing accent) while the second and the fourth one together correspond to another accent set SH. They are separated in the vertical direction. It is obvious that this component has strong correlation with accents.

To illustrate the projection of the four different subsets onto the top two components, we draw each speaker with a point in Figure 3.4. The distribution spans a 2-d speaker space. It can be concluded that the gender and accent are the two main components that constitute the speaker space.

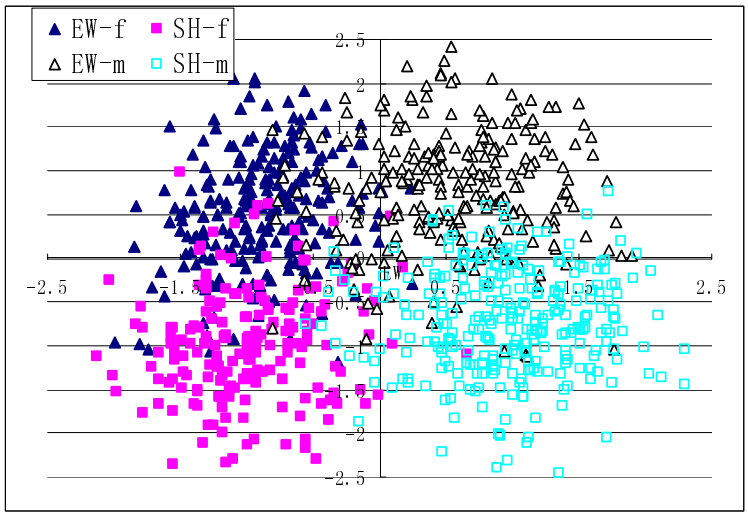

Figure 3.4: Projection of all speakers on the first and second independent components, horizontal direction is the projection on first independent component; vertical direction is projection on second independent component.

To illustrate accurately the performance of ICA, we compute the classification errors on the gender and accent classification through proper choice of projection threshold on each dimension shown in Figure 3.4. There are 60 and 130 errors for gender and accent, respectively. The corresponding error rates are 6.1% and 13.3%.

3.3.4 ICA vs. PCA

When applying PCA and ICA to gender classification on EW corpus, we received the error rate of 13.6% and 8.4% respectively. The results are achieved with the following setups to represent each speaker:

- 6 supporting regression classes;
- Diagonal matrix (~d)
- Static cepstrum and energy (13)

The similar results are achieved with other settings. It is shown that ICA based features yield better classification performance than PCA ones.

Unlike PCA where the components can be ranked according to the eigenvalues, ranking of the positions of the ICA components representing variations in gender and accent can not be done. However, we can always identify them in some way (e.g. from plots). Once they are determined, the projection matrix is fixed.

3.4 Conclusion

In this section, we investigated the variability between speakers through two powerful multivariate statistical analysis methods, PCA and ICA. It is found that strong correlations between gender and accent exist in two ICA components. While strong correlation between gender and the first PCA component is well known, we give the first physical interpretation for the second component: it is strongly related with accent.

We propose to do a proper selection of supporting regression classes, to obtain an efficient speaker representation. This is beneficial for speaker adaptation with limited corpus available.
Through gender classification experiments combined with MLLR and PCA, we concluded that the static and first–order cepstrum and energy carry most information about speakers.

The features extracted by using PCA and ICA analysis can be directly applied to speaker clustering. Further work of its application in speech recognition is undergoing.

4. Accent Modeling through PDA

4.1 Introduction

There are multiple accents in Mandarin. A speech recognizer built for a certain accent often obtains 1.5 ~ 2 times higher error rate when applied to another accent. The errors can be divided into two categories. One type of errors is due to misrecognition of confusable sounds by the recognizer. The other type of errors is those due to the speaker's own pronunciation errors. For example, some speakers are not able to clearly enunciate the difference between /zh/ and /z/. Error analysis shows that the second type of errors constitutes a large proportion of the total errors when a speech recognizer trained on Beijing speakers is applied to speech from Shanghai speakers. A key observation is that speakers belonging to the same accent region have similar tendencies in mispronunciations.

Based on the above fact, an accent modeling technology called 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 utterances and encode these differences into the dictionary, called an accent-specific dictionary. The goal is to estimate the pronunciation differences, mainly consisting of confusion pairs, reliably and correctly. Depending on the amount of the adaptation data, a dynamic dictionary construction process is presented in multiple levels such as phoneme, base syllable and tonal syllable. Both context-dependent and context-independent pronunciation models are also considered. To ensure that the confusion matrices reflect the accent characteristics, both the occurrences of reference observations and the probability of pronunciation variation are taken into account when deciding which transformation pairs should be encoded into the dictionary.

In addition, to verify that pronunciation variation and acoustic deviation are two important but complementary factors affecting the performance of recognizer, maximum likelihood linear regression (MLLR) [11], a well-proven adaptation method in the field of acoustic model was adopted in two modes: separately and combined with PDA.

Compared with [7], which synthesizes the dictionary completely from the adaptation corpus; we augment the process by incorporating obvious pronunciation variations into the accent-specific dictionary with varying weights. As a result, the adaptation corpus that was used to catch the accent characteristics could be comparatively small. Essentially, the entries in the adapted dictionary consist of multiple pronunciations with prior probability that reflect accent variation. In [8], syllable-based context was considered. We extend such context from the syllable to the phone level, even the phone class level. There are several advantages. It can extract the essential variation in continuous speech from a limited corpus. At the same time, it can maintain a detailed description of the impact of articulation of pronunciation variation. Furthermore, tonal changes, as a part of pronunciation variation have also been modeled. In addition, the result we reported has incorporated a language model. In other words, these results could accurately reflect contribution of PDA, MLLR and the combination of two in the dictation application. As we know, a language model could help to recover from some errors due to speakers' pronunciation variation.

Furthermore, most prior work [7][8][10] uses pronunciation variation information to re-

score the N-best hypothesis or lattices resulting from the baseline. However, we developed a one-pass search strategy that unifies all kinds of information, including acoustic model, language model and accent model about pronunciation variation, according to the existing baseline system.

4.2 Accent Modeling With PDA

Many adaptation technologies based on acoustic model parameter re-estimation make assumption that speakers, even in different regions, pronounce words according to a predefined and unified manner. Error analyses across different accent regions tell us that this is a poor assumption. For example, a speaker from Shanghai probably utters /shi/ as /si/ in the canonical dictionary (such as the official published one based on pronunciation of Beijing inhabitants). Therefore, a recognizer trained according to the pronunciation criterion of Beijing cannot recognize accurately a Shanghai speaker given such a pronunciation discrepancy. The aim of PDA is to build a pronunciation dictionary suited to the accent-specific group in terms of a "native" recognizer. Luckily, pronunciation variation between accent groups presents certain clear and fixed tendencies. There exist some distinct transformation pairs at the level of phones or syllables. This provides the premise to carry out accent modeling through PDA. The PDA algorithm can be divided into the following stages:

The first stage is to obtain an accurate syllable level transcription of the accent corpus in terms of the phone set of the standard recognizer. To reflect factual pronunciation deviation, no language model was used here. The transcribed result was aligned with the reference transcription through dynamic programming. After the alignments, error pairs can be identified. Here, we just consider the error pairs due to substitution error since insertion and deletion errors are infrequent in Mandarin because of the strict syllable structure. To ensure that the mapping pairs were estimated reliably and representatively, pairs with few observations were cut off. In addition, pairs with low transformation probability were also eliminated to avoid excessive variations for a certain lexicon items. According to the amount of accent corpus, context dependent or context independent mapping pairs with different transfer probability could be selectively extracted at the level of sub-syllable, base-syllable or tone-syllable.

The next step is to construct a new dictionary that reflects the accent characteristics based on the transformation pairs. We encode these pronunciation transfer pairs into the original canonical lexicon, and finally a new dictionary adapted to a certain accent is constructed. In fact, pronunciation variation is realized through multiple pronunciations with corresponding weights. Each dictionary entry can be a word with multiple syllables or just a single syllable. Of course, all the pronunciation variations' weights corresponding to the same word should be normalized.

The final step is to integrate the adapted dictionary into the recognition or search framework. Much work makes use of PDA through multiple-pass search strategy [8][10]. In other words, prior knowledge about pronunciation transformation was used to re-score the multiple hypotheses or lattice obtained in the original search procedure. In this paper,

we adopt a one-pass search mechanism as in Microsoft Whisper System [9]. Equivalently, the PDA information was utilized at the same time as other information, such as language model and acoustic evaluation. This is illustrated with the following example.

For example: speakers with a Shanghai accent probably uttered “du2-bu4-yi1-shi2” from the canonical dictionary 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) (3).

$$Score(shi2 \mid yi1) = w_{LM} * P_{LM}(shi2 \mid yi1) + w_{AM} * P_{AM}(shi2) + w_{PDA} * P_{PDA}(shi2 \mid shi2) \quad (3)$$

$$Score(shi2(2) \mid yi1) = w_{LM} * P_{LM}(shi2(2) \mid yi1) + w_{AM} * P_{AM}(shi2(2)) + w_{PDA} * P_{PDA}(shi2(2) \mid shi2) = w_{LM} * P_{LM}(shi2 \mid yi1) + w_{AM} * P_{AM}(si2) + w_{PDA} * P_{PDA}(si2 \mid shi2) \quad (4)$$

$$Score(si2 \mid yi1) = w_{LM} * P_{LM}(si2 \mid yi1) + w_{AM} * P_{AM}(si2) + w_{PDA} * P_{PDA}(si2 \mid si2) \quad (5)$$

Where P_{LM}, P_{AM} and P_{PDA} stand for the logarithmic score of Language Model (LM), Acoustic Model (AM) and Pronunciation variation respectively. w_{LM}, w_{AM} and w_{PDA} are the corresponding weight coefficients and adjusted according to experience.

Obviously, the partial path yi1→shi2 (4) has adopted the factual pronunciation (as / $si2$ /) while keeping the ought-to-be LM, e.g. bigram of ($shi2 \mid yi1$), 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 user’s pronunciation error using PDA.

4.3 Experiments and Result

4.3.1 System and Corpus

Our baseline system is an extension of the Microsoft Whisper speech recognition system [9] that focuses on Mandarin characteristics, e.g. pitch and tone have been successfully incorporated [3]. The acoustic model was trained on a database of 100,000 sentences collected from 500 speakers (train_set, male and female half each, here we only use 250 male speakers) coming from Beijing area. The baseline dictionary is based on an official published dictionary that is consistent with the base recognizer. The language model is

tonal syllable trigram with perplexity of 98 on the test corpus. Other data sets are as follows:

- Dictionary Adaptation Set (pda_set): 24 male speakers from Shanghai area, at most 250 sentences or phrases from each speaker;
- Test Set (Test_set) 10 male speakers, 20 utterances from each speaker;
- MLLR adaptation sets (mllr_set): Same speaker set as test sets, at most another 180 sentences from each speaker;
- Accent specific SH model (SH_set): 480 speakers from Shanghai area, at most 250 sentences or phrase from each speaker. (Only 290 male speakers used)

4.3.2 Analysis

2000 sentences from 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 many interesting linguistic phenomena were observed.

Front nasal and back nasal

Final ING and IN are often exchangeable, while ENG are often uttered into EN and not vice versa. This is shown in Table 4.1.

Table 4.1: Front nasal and back nasal mapping pairs of accent speaker in term of standard phone set.

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

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

Because of phonemic diversity, it is hard for Shanghai speakers to utter initial phoneme 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 2. It reveals a strong correlation with phonological observations.

Table 4.2: Syllable mapping pairs of accented speakers in term of standard phone set:

Canonical Pron.	Observed Pron.	Prob. (%)	Canonical Pron.	Observed Pron.	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

4.3.3 Result

In this subsection, we report our result with PDA only, MLLR only and the combination of PDA and MLLR sequentially. To measure the impact of different baseline system on the PDA and MLLR, the performance of accent-dependent SI model and mixed accent groups SI model are also present in both syllable accuracy and character accuracy for LVCSR.

PDA Only

Starting with many kinds of mapping pairs, we first remove pairs with fewer observation and poor variation probability, and encode the remaining pairs into dictionary. Table 4.3 shows the result when we use 37 transformation pairs, mainly consisting of pairs shown in Table 4.1 and Table 4.2.

Table 4.3: Performance of PDA (37 transformation pairs used in PDA).

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

MLLR

To evaluate the acoustic model adaptation performance, we carry out the MLLR experiments. All phones (totally 187) were classified into 65 regression classes. Both diagonal matrix and bias offset were used in the MLLR transformation matrix. Adaptation set size ranging from 10 to 180 utterances for each speaker was tried. Results are shown in the Table 4.4. It is shown that when the number of adaptation utterances reaches 20, relative error reduction is more than 22%.

Table 4.4: Performance of MLLR with different adaptation sentences.

# Adaptation Sentences	0	10	20	30	45	90	180
MLLR	23.18	21.48	17.93	17.59	16.38	15.89	15.50
Error reduction (Based on SI)	--	7.33	22.65	24.12	29.34	31.45	33.13

Combined PDA and MLLR

Based on the assumption that PDA and MLLR can be complementary adaptation technologies from the pronunciation variation and acoustic characteristics respectively, experiment combining MLLR and PDA were carried out. Compared with performance without adaptation at all, 28.4% was achieved (only 30 utterances used for each person). Compared with MLLR alone, a further 5.7% was improved.

Table 4.5: Performance Combined MLLR with PDA.

# Adaptation Sentences	0	10	20	30	45	90	180
+ MLLR + PDA	19.96	21.12	17.5	16.59	15.77	15.22	14.83
Error reduction (Based on SI)	13.9	8.9	24.5	28.4	32.0	34.3	36.0
Error reduction (Based on MLLR)	-	1.7	2.4	5.7	3.7	4.2	4.3

Comparison of Different Models

The following table shows the results of different baseline models or different adaptation techniques on recognition tasks across accent regions. It shows that accent-specific model still outperforms any other combination.

Table 4.6: Syllable error rate with different baseline model or different adaptation technologies (BES means a larger training set including 1500 speakers from both Beijing and Shanghai).

Different Setup	Different Baseline (Syllable Error Rate (%))		
	Train_set	BES	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
+ MLLR + PDA	16.59	14.31	13.52

PDA and MLLR in LVCSR

To investigate the impact of the above strategies on large vocabulary speech recognition, we designed a new series of experiments to be compared with results shown in Table 4.6. A canonical dictionary consisting of up to 50K items and language model of about 120M were used. The result is shown in Table 4.7. Character accuracy is not as significant as syllable accuracy shown in Table 6. It is mainly due to the following two simplifications: Firstly, because of the size limitation of dictionary, only twenty confusion pairs were encoded into pronunciation dictionary. Secondly, no probability is assigned to each pronunciation entry at present. However, we still can infer that PDA is a powerful accent modeling method and is complementary to MLLR.

Table 4.7: Character error rate with different baseline model or different adaptation technologies (BES means a larger training set including 1500 speakers from both Beijing and Shanghai).

Different Setup	Different Baseline (Character Error Rate (%))		
	Train_set	BES	SH_set
Baseline	26.01	21.30	18.26
+ PDA	23.64	20.02	18.41
+ MLLR (30 Utts.)	21.42	18.99	18.51
+ MLLR + PDA	20.69	18.87	18.35
+ MLLR (180 Utts.)	19.02	18.60	17.11

5. Automatic Accent Identification

5.1 Introduction

Speaker variability, such as gender, accent, age, speaking rate, and phones realizations, is one of the main difficulties in speech recognition task. It is shown in [12] that gender and accent are the two most important factors in speaker variability. Usually, gender-dependent model is used to deal with the gender variability problem.

In China, almost every province has its own dialect. When speaking Mandarin, the speaker's dialect greatly affects his/her accent. Some typical accents, such as Beijing, Shanghai, Guangdong and Taiwan, are quite different from each other in acoustic characteristics. Similar to gender variability, a simple method to deal with accent problem is to build multiple models of smaller accent variances, and then use a model selector for the adaptation. Cross-accents experiments in Section 2 show that performance of accent-independent system is generally 30% worse than that of accent-dependent one. Thus it is meaningful to develop an accent identification method with acceptable error rate.

Current accent identification research focuses on foreign accent problem. That is, identifying non-native accents. Teixeira et al. [13] proposed a Hidden Markov Model (HMM) based system to identify English with 6 foreign accents: Danish, German, British, Spanish, Italian and Portuguese. A context independent HMM was used since the corpus consisted of isolated words only, which is not always the case in applications. Hansen and Arslan [14] also built HMM to classify foreign accent of American English. They analyzed some prosodic features' impact on classification performance and concluded that carefully selected prosodic features would improve the classification accuracy. Instead of phoneme-based HMM, Fung and Liu [15] used phoneme-class HMMs to differentiate Cantonese English from native English. Berkling et al. [16] added English syllable structure knowledge to help recognize 3 accented speaker groups of Australian English.

Although foreign accent identification is extensively explored, little has been done to domestic one, to the best of our knowledge. Actually, domestic accent identification is more challenging: 1) Some linguistic knowledge, such as syllable structure used in [16], is of little use since people seldom make such mistakes in their mother language; 2) Difference among domestic speakers is relatively smaller than that among foreign speakers. In our work, we want to identify different accent types spoken by people with the same mother language.

Most of current accent identification systems, as mentioned above, are built based on the HMM framework, while some investigated accent specific features to improve the performance. 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 economic. We just want to know which accent type the given utterances belong to. Furthermore, HMM training is a supervised one: it needs phone transcriptions. The transcriptions are either manually labeled, or obtained from a speaker independent model, in which the word error rate will certainly degrade the identification performance.

In this section, we propose a GMM based method for the identification of domestic speaker accent. Four typical Mandarin accent types are explored: Beijing, Shanghai, Guangdong and Taiwan. Since phoneme or phoneme class information are out of our concern, we just model accent characteristics of speech signals. GMM training is an unsupervised one: no transcriptions are needed. We train two GMMs for each accent: one for male, the other for female, since gender is the greatest speaker variability. Given test utterances, the speaker's gender and accent can be identified at the same time, compared with the two-stage method in [13]. The commonly used feature in speech recognition systems, MFCC, is adopted to train the GMMs. The relationship between GMM parameter and recognition accuracy is examined. We also investigate how many utterances per speaker are sufficient to reliably recognize his/her accent. We randomly select N utterances from each test speaker and averaged their log-likelihood in each GMM. It is hoped that the more the averaged utterances, the more robust the identification results. Experiments show that with 4 test utterances per speaker, about 11.7% and 15.5% error rate in accent classification is achieved for female and male, respectively. Finally, we show the correlations among accents, and give some

explanations.

5.2 Multi-Accent Mandarin Corpus

The multi-accent Mandarin corpus, consisting of 1,440 speakers, is part of 7 corpora for speech recognition research collected by Microsoft Research China. There are 4 accents: Beijing (BJ, including 3 channels: BJ, EW, FL), Shanghai (SH, including 2 channels: SH, JD), Guangdong (GD) and Taiwan (TW). All waveforms were recorded at a sampling rate of 16 kHz, except that the TW ones were 22 kHz. Most of the data were from students and staff at universities in Beijing, Shanghai, Guangdong and Taiwan, with ages varying from 18 to 40. In training corpus, there are 150 female and 150 male speakers of each accent, with 2 utterances per speaker. In test corpus, there are 30 female and 30 male speakers of each accent, with 50 utterances per speaker. Most of the utterances last about 3-5 seconds each, forming about 16 hours' speech data of the whole corpus. There is no overlap between training and test corpus. That is, all the 1,440 speakers are different. The speaker distribution of the multi-accent Mandarin corpus is listed in Table 5.1.

Table 5.1: Speaker Distribution of Corpus.

Accent	Channel	Gender	Training Corpus		Test Corpus	
BJ	BJ	F	50	300	10	60
		M	50		10	
	EW	F	50		10	
		M	50		10	
	FL	F	50		10	
		M	50		10	
SH	SH	F	75	300	15	60
		M	75		15	
	JD	F	75		15	
		M	75		15	
GD	GD	F	150	300	30	60
		M	150		30	
TW	TW	F	150	300	30	60
		M	150		30	
ALL			1,200		240	

5.3 Accent Identification System

Since gender and accent are important factors of speaker variability, the probability density functions of distorted features caused by different gender and accent are different. As a result, we can use a set of GMMs to estimate the probability that the observed utterance is come from a particular gender and accent.

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

likelihood function

$$\log\prod_{t=1}^{T} p(x(t)\mid\Lambda_k)=\sum_{t=1}^{T}\log p(x(t)\mid\Lambda_k), k=1,...,M, \qquad (6)$$

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 [17]. During identification, an utterance is fed to all the GMMs. The most likely gender and accent type is identified according to

$$\hat{k}=\arg\max_{k=1}^{M}\sum_{t=1}^{T}\log p(x(t)\mid\Lambda_k). \qquad (7)$$

5.4 Experiments

5.4.1 Experiments Setup

As described in Section 5.2, there are 8 subsets (accent plus gender) in the training corpora. In each subset, 2 utterances per speaker, altogether 300 utterances per subset, are used to train the GMMs. Since the 300 utterances in a subset are from 150 speakers with different ages, speaking rates and even recording channels, speaker variability caused by these factors is averaged. Thus we hope to represent effectively the specific gender and accent by this means. The speech data is pre-emphasized with $H(z)=1-0.97z^{-1}$, windowed to 25-ms frames with 10-ms frame shift, and parameterized into 39 order MFCCs, consisting of 12 cepstral coefficients, energy, and their first and second order differences. Cepstral mean subtraction is performed within each utterance to remove the effect of channels. When training GMMs, their parameters are initialized and reestimated once. Data preparation and training procedures are performed using the HTK 3.0 toolkit [19]. In the first experiment, we investigate the relation between the number of components in GMM and the identification accuracy.

50 utterances of each speaker are used for test. In the second experiment, we study how the number of utterances affects the performance of our method. For each test speaker, we randomly select N (N<=50) utterances and average their log-likelihood in each subset. The test speaker is classified into the subset with the largest averaged log-likelihood. The random selection is repeated for 10 times. Thus 2400 tests are performed in each experiment. This will guarantee to achieve reliable results.

5.4.2 Number of Components in GMM

In this experiment, we examine the relationship between the number of components in GMMs and the identification accuracy.

Since our problem is to classify the unknown utterances to a specific subset, and the eight subsets are labeled with gender and accent, our method can identify the speaker's gender and accent at the same time. When calculating the error rate of gender, we just concern with speakers whose identified gender is different with the labeled one. Similarly, when calculating the error rate of accent, we just concern with speakers whose identified accent is error.

Table 5.2 and Figure 5.1 show the gender and accent identification error rate respectively, varying the number of components in GMMs. Here also listed the relative error reduction as increasing the number of components.

Table 5.2: Gender Identification Error Rate(Relative error reduction is calculated when regarding GMM with 8 components as the baseline).

# of Components	8	16	32	64
Error Rate (%)	8.5	4.5	3.4	3.0
Rel. Error Reduction (%)	-	47.1	60.0	64.7

Table 5.2 shows that the gender identification error rate decreases significantly when components increase from 8 to 32. However, only small improvement is gained by using 64 components compared with 32 ones. It can be concluded that GMM with 32 components is capable of effectively modeling gender variability of speech signals.

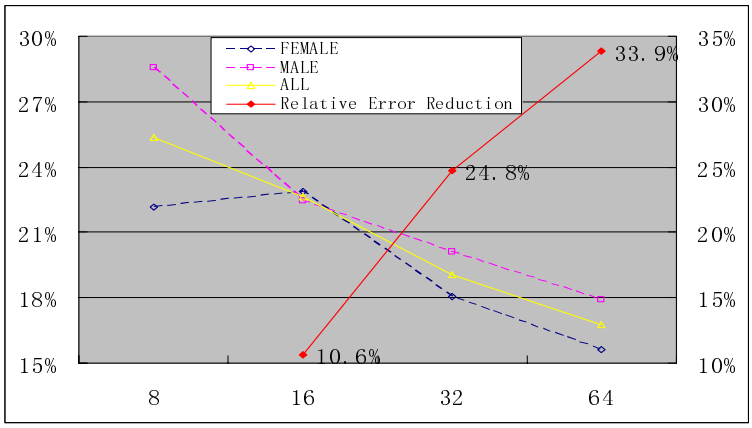

Figure 5.1: Accent identification error rate with different number of components. X axis is the number of components in GMMs. The left Y axis is the identification error rate; the right Y axis is the relative error reduction to 8 components, when regarding GMM with 8 components as the baseline. "All" means error rate averaged between female and male.

Figure 5.1 shows the similar trend with Table 5.2. It is clear that the number of components in GMMs greatly affects the accent identification performance. Different to the gender experiment, in accent, GMMs with 64 components still gain some improvement over 32 ones (Error rate decreases from 19.1% to 16.8%). Since the accent variability in speech signals is more complicated and not as significant as gender, 64 components are better while describing the detail variances among accent types.

However, it is well known that to train a GMM with more components is much more time-consuming and requires more training data to obtain reliable estimation of the parameters. Concerning the trade-off between accuracy and costs, using GMMs with 32 components is a good choice.

5.4.3 Number of Utterances per Speaker

Sometimes it is hard even for linguistic experts to tell a specific accent type given only one utterance. Thus making use of more than one utterance in accent identification is acceptable in most applications. We want to know the robustness of the method: how many utterances are sufficient to reliably classify accent types.

In this experiment, we randomly select N (N<=50) utterances for each test speaker and average their log-likelihood in each GMM. The test speaker is classified into the subset with the largest averaged log-likelihood. The random selection is repeated for 10 times to guarantee achieving reliable results. According to Section 5.3.2, 32 components for each GMM are used.

Table 5.3 and Figure 5.2 show the gender and accent identification error rate respectively, varying the number of utterances. When averaging the log-likelihood of all 50 utterances of a speaker, it is no need to perform random selection.

Table 5.3: Gender Identification Error Rate (Relative error reduction is calculated when regarding 1 utterance as the baseline).

# 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

Table 5.3 shows that it is more reliable to tell a speaker's gender by using more utterances. When the number of utterances increases from 1 to 4, the gender identification accuracy improves greatly. Still considerable improvement is observed when using more than 10 utterances. However, in some applications, it is not applicable to collect so much data just to identify the speaker's gender. Also, the results of 3~5 utterances are good enough in most situations.

It is clear from Figure 5.2 that increasing the number of utterances improves identification performance. This is consistent with our idea that more utterances of a speaker, thus more information, help recognize his/her accent better. Considering the trade-off between accuracy and costs, using 3~5 utterances is a good choice, with error rate 13.6%-13.2%.

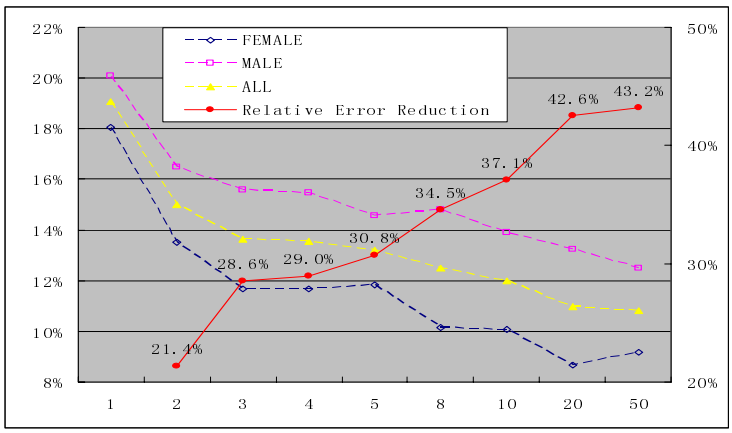

Figure 5.2: Accent identification error rate with different number of utterances. X axis is the number of utterances for averaging. The left Y axis is the identification error rate; the right Y axis is the relative error reduction, when regarding 1 utterance as the baseline. "All" means error rate averaged between female and male.

5.4.4 Discussions on Inter-Gender and Inter-Accent Results

It can be noticed from Figure 5.1 and Figure 5.2 that the accent identification results are different between male and female. In experiments we also discovered different pattern of identification accuracy among 4 accent types. In this subsection, we will try to give some explanations.

We select one experiment in Section 5.4.3 as an example to illustrate the two problems. Here GMMs are built with 32 components. 4 utterances of each speaker are used to calculate the averaged log-likelihood to recognize his/her accent. The inter-gender result is listed in Table 5.4. Table 5.5 shows the recognition accuracy of the 4 accents.

Table 5.4: Inter-Gender Accent Identification Result.

Error Rate (%)	BJ	SH	GD	TW	ALL Accents
Female	17.3	11.4	15.2	2.7	11.7
Male	27.7	26.3	7.6	0.3	15.5

We can see from Table 5.4 that Beijing (BJ) and Shanghai (SH) female speakers are much better recognized than corresponding male speakers, which causes the overall better performance for female. This is consistent with speech recognition results. Experiments in Section 2 show better recognition accuracy for female than for male in Beijing and Shanghai, while reverse result for Guangdong and Taiwan.

Table 5.5 shows clearly different performance among accents. We will give some discussions below.

Table 5.5: Accents identification confusion matrices (Including four accents like Beijing, Shanghai, Guangdong and Taiwan).

Recognized As	Testing Utterances From			
	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 most likely to be recognized to each other, except to themselves. In fact, Shanghai and Guangdong both belong to southern language tree in phonology and share some common characteristics. For example, they do not differentiate front nasal and back nasal.

- The excellent result of Taiwan speakers may lie in two reasons. Firstly, as Taiwan civilians communicate with the Mainland relatively infrequently and their language environment is unique, their speech style is quite different from that of the Mainland people. Secondly, limited by the recording condition, there is a certain portion of noise in the waveform of Taiwan corpus (both training and test), which makes them more special.

- The reason of relatively low accuracy of Beijing possibly lies in its corpus's channel variations. It is shown in Table 5.1 there are 3 channels in Beijing corpus. Greater variations lead to a more general model, which is not so specific for the accent and may degrade the performance.

- Channel effect may be a considerable factor to the GMM based accent identification system. From Beijing, Shanghai and Guangdong, accuracy increases when the number of channels decreases. Further work is needed to solve this problem.

6. Conclusion and Discussions

Accent is one of the main factors that cause speaker variances and a very serious problem that affects speech recognition performance. We have explored such problem in two directions:

- Model adaptation. Pronunciation dictionary adaptation method is proposed to catch the pronunciation variation between speakers based on standard dictionary and the accented speakers. In addition pronunciation level adjustments, we also tried model level adaptation such as MLLR and integration of these two methods. Pronunciation adaptation can cover most dominant variation among accents group in phonology level, while speaker adaptation can tract the detailed changes for

specific speaker such pronunciation style in acoustic level. Result shows that they are complimentary.

- Building accent specific model and automatic accent identification. In case we have enough corpus for each accent, we can build more specific model with little speaker variances. In the report, we propose a GMM based automatic accent detection method. Compared with HMM based identification methods. It has the following advantages. Firstly, it is not necessary to know the transcription in advance. In other word, it is text independent. Secondly, because the parameter need to be estimated is far less, it greatly reduced the enrollment burden of the users. Lastly, it is very efficient to identify the accent type of new comers. In addition, the method can be extended any more detailed speaker subset of certain characteristics, such as more detailed classification about speakers.

There two methods can be adopted in different case according to the mount of available corpus. When large amount of corpora for different accents can be obtained, we get classify the speaker through GMM-based automatic accent identification strategies proposed in Section 5 into different speaker subsets and train accent specific model respectively. Otherwise, we can extract the main pronunciation variations between accent groups and standard speakers through PDA with a certain amount of accented utterances.

Furthermore, we make a thorough investigation among speaker variability, especially focusing on gender and accent. In the process, we proposed MLLR transformations-based speaker representation and introduced supporting regression class concept. Finally, we have given the physical interpretation of accent and gender. That is: The two factors have strong correlation with the first two independent components, which bridge the gap between low-level speech events, such as features, and the high-end speaker characteristics: accent and gender.

7. References

[1] A. Hyvarinen and E. Oja, “Independent component analysis: algorithms and application,” *Neural Networks*, vol. 13, pp.411-430, 2000.

[2] H. Hotellings, “Analysis of a complex of statistical variables into principle components,” *J. Educ. Psychol.*, vol. 24, pp.417-441, 498-520, 1933.

[3] E. Chang, J. L. Zhou, C. Huang, S. Di, K. F. Lee, “Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones,” In *Proc. of ICSLP’2000*, Beijing, Oct. 2000.

[4] N. Malayath, H. Hermansky, and A. Kain , “Towards decomposing the sources of variability in speech,” in *Proc. Eurospeech’97*, vol. 1, pp. 497-500, Sept. 1997.

[5] R. Kuhn, J. C. Junqua, P. Nguyen and N. Niedzielski, “Rapid Speaker Adaptation in Eigenvoice Space,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, n6, Nov. 2000.

[6] Z. H. Hu, “Understanding and adapting to speaker variability using correlation-based principal component analysis”, *Dissertation of OGI*. Oct. 10, 1999.

[7] J. J. Humphries and P.C. Woodland, "The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training," in *Proc. ICASSP'98*, vol.1, pp. 317-320, Seattle, USA, 1998.

[8] M. K. Liu, B. Xu, T. Y. Huang, Y. G. Deng, C. R. Li, "Mandarin Accent Adaptation Based on Context-Independent/Context-Dependent Pronunciation Modeling," in *Proc. ICASSP'2000*, vol.2, pp. 1025-1028, , Turkey, 2000.

[9] X. D. Huang, A. Acero, F. Alleva, M. Y. Hwang, L Jiang, M. Mahajan, "Microsoft Windows highly intelligent speech recognizer: Whisper," in *Proc. ICASSP'95*, vol. 1, pp. 93-96, 1995.

[10] M. D. Riley and A. Ljolje, "Automatic Generation of Detailed Pronunciation Lexicon," Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer. 1995.

[11] C. J. Leggetter, P. C. Woodland, "Maximum likely-hood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, n2, pp. 171-185, April, 1995.

[12] C. Huang, T. Chen, S. Li, E. Chang and J.L. Zhou, "Analysis of Speaker Variability," in *Proc. Eurospeech'2001*, vol.2, pp.1377-1380, 2001.

[13] C. Teixeira, I. Trancoso and A. Serralheiro, "Accent Identification," in *Proc. ICSLP'96*, vol.3, pp. 1784-1787, 1996.

[14] J.H.L. Hansen and L.M. Arslan, "Foreign Accent Classification Using Source Generator Based Prosodic Features," in *Proc. ICASSP'95*, vol.1, pp. 836-839, 1995.

[15] P. Fung and W.K. Liu, "Fast Accent Identification and Accented Speech Recognition," in *Proc. ICASSP'99*, vol.1, pp. 221-224, 1999.

[16] K. Berkling, M. Zissman, J. Vonwiller and C. Cleirigh, "Improving Accent Identification Through Knowledge of English Syllable Structure," in *Proc. ICSLP'98*, vol.2, pp. 89-92, 1998.

[17] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, vol.39, pp. 1-38, 1977.

[18] C. Huang, E. Chang, J.L. Zhou, K.F. Lee, "Accent Modeling Based On Pronunciation Dictionary Adaptation For Large Vocabulary Mandarin Speech Recognition", Vol.3 pp.818-821, ICSLP'2000, Oct. Beijing.

[19] http://htk.eng.cam.ac.uk.