Chinese Dialect-Based Speaker Classification and Pronunciation Assessment Using Structural Representation of Speech

音声構造表象を用いた中国語方言に基づく話者分類と発音評価

Xuebin MA 47-077312

Supervisor: Prof. Nobuaki MINEMATSU

Frontier Informatics
Graduate School of Frontier Sciences
The University of Tokyo

Doctor's thesis

10.06.2010

Declaration

I hereby declare that I prepared the Ph.D. thesis “Chinese Dialect-Based Speaker Classification and Pronunciation Assessment Using Structural Representation of Speech” on my own and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Nobuaki Minematsu for his always insightful suggestions and expert guidance. As a great educationist and an excellent researcher, his unwavering commitment to his students and constant demand for excellence really helped me bring this work to fruition. It has been a privilege and memorable experience to work with him.

Secondly, I would also like to thank Professor Keikichi Hirose and all the other members of "Hirose-Minematsu Lab." for providing an excellent and inspiring working atmosphere. Dr. Yu Qiao gave me many constructive suggestions for my study and also helped me a lot in the everyday life. Ms. Aki Kunikoshi had been my daily life tutor for one year and helped me be used to the life here. She taught me many things such as Japanese, Japanese manners and she also became a good friend of mine. Mr. Daisuke Saito was always willing to help others and he helped me a lot to use the resources of the Lab, fix some programming problems and so on. We also had a lot of fun drinking beer together. About the other members of this lab, although I cannot list all their names here, I am very grateful to them for leaving me so many memorable experiences.

I want to express my special thanks to Mr. Akira Nemoto, Nankai University, for helping me collecting many dialect data in China. He also spent many private times helping me label the data. I also want to thank Dr. Guoping Hu and some other friends of the University of Science and Technology of China for sharing me some experimental data. I also want to express my special thanks to Professor Aijun Li and Dr. Ruiyuan Xu, the Institute of Linguistics, the Chinese

Academy of Social Sciences, for helping me finish the most challenging work of building a new database with constant speaker identity.

I would also thank my colleagues from the Chinese Student and Scholar Association of the University of Tokyo, my friends from the basketball club, and some other acquaintances in this university. Thank you for giving me a lot of great memorable experiences.

Finally, I owe special gratitude to my parents and my girlfriend, who will be my wife this year, for their continuous and unconditional support, understanding, and love.

Abstract

In modern speech processing technologies, segmental features of speech are usually represented acoustically by spectrum, which contains not only linguistic information but also extra-linguistic information corresponding to age, gender, speaker, microphone, and so on. If one wants to classify speakers using their utterances purely based on their dialects, only the dialectal differences should be focused on and the extra-linguistic features should be removed or canceled. In fact, for the problems of automatic speech recognition, very similar problems are raised where the linguistic features of speech invariant or robust to extra-linguistic factors are desired. Therefore, a method to build so-called speaker-independent models is studied by collecting the data of many speakers trying to cover all the extra-linguistic features. About some linguistic studies, in order to compare the vowel realizations of different speakers in linguistic and sociolinguistic meaningful ways, normalization techniques are used to capture the differences. However, these methods may not work well in the problem of Chinese dialect-based speaker classification. For this problem, the linguistic features invariant to extra-linguistic factors should be extracted from the dialect utterances of individual speakers.

In our previous works, a structural representation of speech is proposed to extract the speech contrasts or dynamics by removing extra-linguistic features from speech and it is already applied to speech recognition, speech synthesis and helping Japanese learning English. In my study, the structural method is further applied to Chinese dialect pronunciations representation and dialect-based speaker classification is achieved by building comparable dialect structures to ex-

tract the speaker-invariant purely linguistic features from Chinese dialects. At the beginning, based on the phonological features of Chinese dialects, utterances of syllable units (characters) are proposed as the reading materials to built pronunciation structures. Then several different lists of Chinese written characters, which are original proposed by Chinese dialectologists to check the dialect pronunciation of different speakers, are adopted as the reading materials to built dialect-sensitive comparable dialect pronunciation structures. After that, using the dialectal utterances of the reading materials, dialect pronunciation structure is built for every speaker by calculating the Bhattacharyya distances between the distributions of any pair of his/her utterances. Because Bhattacharyya distance is invariant to affine transformations and extra-linguistic features perform as affine transformations in spectral space, the built dialect pronunciation structure is invariant to extra-linguistic features in speech. Therefore, speaker-invariant dialect-based speaker classification can be achieved by building the dialect pronunciation structures for the speakers and calculating the distances between their pronunciation structures.

In order to verify my proposal, several different classification experiments are carried out. At the beginning, a dialect-based speaker classification experiment is carried out. Because publicly available Chinese dialect corpora cover only two or three dialects and cannot be used for this problem, a new database of Chinese dialects is built and the dialect data of 17 speakers are recorded. Then all the data are labeled manually and the syllables are cut and converted into distributions. After that, for every speaker, the BDs between any pair of his/her utterances are calculated and the pronunciation structure is built. Then speaker classification experiment is carried out by calculating the distances between their pronunciation structures. The result shows that the speakers are well classified by their dialects and the result is independent to extra-linguistic features such as the gender and age of the speakers.

After that, this structural method is verified by a sub-dialect based speaker classification experiment. At the beginning, a new database of sub-dialects is built and the sub-dialect data of 16 speakers from 4 sub-dialects regions of Mandarin and the data are recorded. Then using the same method as last experiment, sub-dialect pronunciation structures are built and these speakers are classified by calculating the distances between their pronunciation structures. By the result, it is found that the speakers from the same dialect cities are all clustered together and the speakers from the same sub-dialect regions are also mainly classified near to each other, except one exception that 4 speakers from ZhongYuan sub-dialect regions are classified to two different sub-trees. Several possible reasons for it are discussed: these speakers are also graduate students in Tianjin and their sub-dialects may be affected by the sub-dialect there to different degrees; the traditional linguistic classification of these sub-dialects are carried out based on several different features of the whole syllable but our method of structural classification is only focusing on the acoustic features of the finals. Anyway, neither of these possible reasons can be proved. So a new evaluation method is proposed to prove that the dialect-based speaker classification using our structural method is not affected by the features of the speakers.

In order to prove that our method can classify speakers by extracting the speaker-invariant linguistic features no matter which kind of dialect are they speaking, new comparison experiments are designed with original dialect data and mimicked dialect data with minimum speaker differences. For these experiments, I carried out some new recordings in China and the data of speakers from 10 sub-dialects of 5 dialect regions were recorded. Then every utterance of this data set is linguistically mimicked by an expert of Chinese dialects and a new data set with fixed speaker identity (minimum speaker differences) is built. After that, using the original and mimicked data separately, dialect-based speaker classification experiments are carried out. It is found that the two results are almost the same as each other, although one is obtained using the dialect data spoken by different speakers and the other is obtained using the dialect data with fixed speaker identity. It means that our method of classify speakers based on their dialects using structural method is really invariant to speakers.

Also, our method of structural pronunciation comparison is compared with conventional spectral comparison using data sets with maximum speaker differences. At the beginning, corresponding to the original and mimicked dialect data used above, new data are converted just like they are pronounced by a very tall speaker and a very short speaker and new data sets with maximum speaker differences are built. Then using these data, classification experiments based on spectral comparisons are carried out. The results show that the classifications are affected greatly by the speaker features. After that, these speakers are classified using our structural method and the results show that they are well classified by their dialects and it is not affected by the speaker differences at all. So our method is proved again that it can classify speakers based on their dialects by extracting the purely linguistic features and the result is not affected by the speaker features like the conventional spectral comparison.

Further, the structural method is applied to estimating the utterance similarity orders between two speakers. Using the dialect data of 2 Min speakers of different genders and the data of 2 standard Mandarin speakers of different genders, experiments are carried out to estimate the utterance similarity orders among them using our structural method. The results show that very similar similarity orders are obtained for the dialect speakers from the same dialect regions and the results are robust to the genders of the speakers. Also, this structural method is applied to pronunciation assessment of accented Mandarin. At the beginning, the accented Mandarin pronunciation structures are built and compared with the pronunciation structures of standard Mandarin. Then a structural score is obtained for every utterance.

After that, these utterances are evaluated manually and the manual evaluated sores are compared with the structural scores. Meanwhile, the data are recognized by a new built Mandarin recognizer and the results are compared with the above two scores. However, the correlation coefficients between these scores are not satisfactory, although some correlations can be found by the results. Therefore, sub-structures are built to assess the accented Mandarin pronunciations. By adding or deleting utterances to built sub-structures, the pronunciations of accented Mandarin speakers are compared with standard Mandarin speakers and the best correlation coefficient is obtained at about 0.4.

Through the above works I have done, it is proved that the structural pronunciation representation can extract the speaker-invariant purely speaker features and classify Chinese dialect speakers based on their dialects. Then we are planning to apply this approach to drawing a new Chinese dialect atlas by calculating the acoustic distances among Chinese dialects, and this result can be further applied to speech processing of multi-dialects. Furthermore, if more data of standard Mandarin pronunciation and well labeled accented Mandarin pronunciation are obtained, I also want to continue the study of pronunciation assessment of accented Mandarin using sub-structure method.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of this study

Nowadays, many researchers are focusing on the studies of dialects and accented languages using modern speech processing techniques. According to the targets, their studies can be classified into two kinds: some linguists are studying dialects based on the acoustic features trying to find some new features of these dialects or some new relationships among them [1, 2, 3]; On the other hand, some studies are done for the target of developing new applications with modern speech technologies, such as dialect identification [4, 5, 6], dialect recognition [7] and so on. However, no matter which kind of studies it is, a problem must be faced that speech contains not only linguistic information but also extra-linguistic information such as the age, gender, speaker and so on.

In fact, in modern speech processing techniques, segmental features of speech are usually represented by spectrum, which contains not only linguistic features but also extra-linguistic features corresponding to age, gender, speaker and so on. But for the problems like dialect-based speaker classification, only the linguistic features are needed and the extra-linguistic features should be canceled or removed. Therefore, in conventional speech processing framework, so-called speaker-independent acoustic models are often built by collecting the data of many different speakers trying to cover all the different features of speakers [8]. Then for the processing of multi-dialects, such speaker-independent models are always built for every dialect using the data of many speakers from this dialect region [9]. About the studies of some sociolinguists and dialectologists, in order to compare the vowel realizations of different speakers in linguistic way, different vowel normalization techniques have been proposed to normalize the distortion caused by the physiological differences among speakers [10, 11, 12]. However, none of these methods can work in our problem of Chinese dialect-based speaker classification. For this problem, the linguistic features invariant to extra-linguistic factors should be extracted from the dialect utterances of individual speakers.

Generally speaking, dialects describe intra-language variety and each of which is used by a particular group of that language's speakers [13]. There are always some phonetic, grammatical and lexical differences to different degrees among dialects. About Chinese dialects, as many of them are mutually unintelligible to each other, some linguists take them as a language family [14, 15, 16]. However, because of many historical and sociological reasons and the following criterion for distinguishing between languages and dialects: "A language is a dialect with an army and navy" [17], we take Chinese dialects as a language and its dialect in this thesis.

In China, there are hundreds kinds of dialects and they are traditionally classified into several major dialect regions [18]. Furthermore, most of these major dialect regions also have many different sub-dialects and sub-sub-dialects. Anyway, all these dialects are developed from the same root and they have inherited a lot of common features. For example, they share the same written characters and phonological structures, every character is pronounced as a mono-syllable which is always composed by an initial, a final and a tone. However, there are still many differences among these dialects in varying degrees grammatically, lexically, phonologically and phonetically [19, 20]. Therefore, people from different major dialect regions always cannot communicate orally. And sometimes, even for the people from adjacent cities, their dialects are quite different and they have difficulties in oral communication. Since 1956, standard Mandarin has been popularized all over the country as the official language and almost every dialect speaker began to learn Mandarin just like learning a second language. However, affected by their native dialects, many of them speak Mandarin with regional accents. On the other hand, affected by the popularization of Mandarin and people are moving across different dialect regions, many dialects are also changing and losing some of their special features. Strictly speaking, every speaker has his/her own dialect not only because speakers of the same dialect are often speakers of different sub-dialects but also because the dialect of this speaker may already changed affected by other dialects or Mandarin.

Considering the complicated situation of Chinese dialects, the method of training different speaker-independent but dialect-dependent models may not work in Chinese dialect-based speaker classification. Because if we want to apply this approach to dialect-based speaker classification, several dozens of dialect or sub-dialects models should be built after the data of many speakers from the same dialect regions are collected. It will be a very challenging work and this approach will conflict with the goal of finding the intra-dialect relations among speakers, because two Chinese speakers of the same dialect may be speakers from different sub-dialect regions. About the linguistic studies of extracting linguistic features using methods like vowel normalization, the dialects of the speakers and the phonological features of these dialects are always needed in advance. In this problem of dialect-based speaker classification, we don't know any information about the dialects of the speakers. Therefore, these linguistic approaches cannot be used in our study either.

In our previous works, a structural representation of speech was proposed to extract the speaker-invariant speech contrasts or dynamics [21, 22]. After the pronunciation structure is built using the interrelations among speech events of every speaker, it can extract the linguistic features by removing the extra-linguistic features and irrelevant acoustic features from speech. Using this approach, Speaker-Independent Automatic Speech Recognition (SI-ASR) was achieved only using a small number of training speakers, where explicit adaptation or normalization was not needed [23, 24]. After that, this approach was applied to helping Japanese learning English [26] and speech synthesis [25], and satisfactory results were obtained.

1.2 Objectives of this study

In my study, in order to solve the problem of classifying Chinese dialect speakers based on their dialects, structural representation of dialect pronunciations are proposed to extract the purely linguistic features of their dialect by canceling the extra-linguistic features. Using the dialect utterances of every speaker, his/her dialect pronunciation structure is built. Then the purely dialectal features can be extracted and it is invariant to extra-linguistic features such as age, gender, speaker and so on. After that, in order to verify this proposal, dialect and sub-dialect-based speaker classification experiment is carried out separately by calculating the distance between any pair of structures and the results are supposed to be invariant to extra-linguistic features. Also, in order to prove the speaker-invariance of this method, two classification experiments using original dialect data and linguistically mimicked data with minimum speaker differences are carried out. At last, corresponding to these original and mimicked data, data sets with maximum speaker differences are built and our structural method is compared with the conventional method based on spectral comparison.

Besides dialect-based speaker classification, the structural representation of dialect pronunciation is also applied to calculating the utterance similarity between the pronunciations of two speakers. By comparing the pronunciation of speakers from the same dialect regions with standard Mandarin, very similar similarity orders should be found. So this method can be applied to finding the common pronunciation features of the speakers who are from the same dialect region. Also, by comparing the pronunciation of accented Mandarin with the pronunciation of standard Mandarin, the accented Mandarin pronunciation can be assessed and the result is irrelevant to the extra-linguistic features in speech.

1.3 Organization of this paper

The rest of this thesis is organized as follows: First, some fundamentals of Chinese dialects, current situation of Chinese spoken language and their new development trend are introduced in Chapter 2. In Chapter 3, some related works are introduced together with some knowledge of conventional dialect processing systems. In Chapter 4, the method of building comparable dialect pronunciation structures and calculating the distance between two structures is introduced. After that, two speaker classification experiments based on Chinese dialects and subdialects of Mandarin are presented separately in Chapter 5 and Chapter 6. In

Chapter 7, the speaker-invariant feature of the structural method is proved by comparison experiments using original dialect data and linguistically mimicked data. In Chapter 8, the structural method is compared with the conventional method based on spectral comparison using data with maximum speaker differences which are created using high-quality voice morphing techniques and sound as if they are produced by the same speaker but with a much longer or shorter vocal tract. In Chapter 9, the structural method is verified whether it can work well in estimating the utterances similarity and pronunciation assessment of accented Mandarin by comparing two pronunciation structures. At last, conclusion of my works is given and the future works are introduced.

Chapter 2

Fundamentals and current situation of Chinese dialects

2.1 Introduction

In this chapter, some fundamentals of Chinese dialects and the current situation of spoken Chinese are introduced. At the beginning, the classification of Chinese dialects is discussed. Then the common and different features among Chinese dialects are presented. After that, the current situation of Chinese spoken languages is introduced. At last, the new development trend of Chinese dialects is presented and this chapter is concluded.

2.2 Fundamentals of Chinese dialects

2.2.1 Classification of Chinese dialects

Generally speaking, dialects describe intra-language variety and each of which is used by a particular group of that language's speakers. There are always some phonetic, grammatical and lexical differences to different degrees among dialects. About Chinese dialects, as many of them are mutually unintelligible to each other, some linguists take Chinese dialects as a language family and its subdivision language [14, 15, 16]. However, because of many historical and sociological reasons and the following criterion for distinguishing between languages and dialects: "A

Table 2.1: Chinese dialects

Group	Speakers (in millions)	Location (Provinces)	Representative sub-dialects
Mandarin	662.2	north of YangZi rivers, and south west provinces	Beijing, Tianjin, Ruicheng
Wu	69.8	south Jiangsu, Zhejiang, south-east Anhui	Shanghai, Suzhou, Danyang
Gan	31.3	Jiangxi, east Hunan	Nanchang
Xiang	30.9	Hunan	ChangSha
Min	55.1	Fujian, Taiwan, east Guandong, Hainan	Fuzhou, Xiamen, Taiwanese
Yue	40.2	Guangdong, east Guangxi	Cantonese, Taishan
Hakka	35.0	south Jiangxi, west Fujian, east Guandong	Meixian, Changting, Pingdong
Jin	45.7	Shanxi, north Shaanxi, west Heberi	Pingyao, Changzhi
Hui	3.1	south-east Anhui, west Zhejiang	Tunxi
PingHua	2.0	south Guangxi	Nanning

language is a dialect with an army and navy" [17], Chinese dialects will be taken as a language and its dialect in this thesis.

In China, there are hundreds kinds of dialects and they can be classified into different groups according to different criteria. According to traditional classification [19, 20], Chinese dialects are classified into 7 major regional groups, i.e., Guanhua(Mandarin), Wu, Xiang, Gan, Keijia(Hakka), Yue(Cantonese) and Min. Meanwhile, some linguists claim that there should be 9 dialect groups (Jin and Hui are added) or more [27, 28]. For example, in [28], Language Atlas of China, Chinese dialects are divided into ten groups as shown in Table 2.1. In this thesis, the traditional classification of Chinese dialects is accepted and the names of these dialect regions are also the same.

In addition, most of Chinese major dialects have many sub-dialects and sub-sub-dialects. For example, Mandarin, spoken by roughly 65 percent of the entire population of China, has 8 sub-dialects all over the country and they can be divided into 42 sub-sub-dialects [18]. For the people from different sub-dialects or sub-sub-dialects regions of Mandarin, most of them can communicate orally. However, at the southeast of China, several major dialects are located there and each of them has many sub-dialects and sub-sub-dialects, people from two adjacent cities may cannot communicate orally because they may speak different dialects or sub-dialects.

2.2.2 Common features among Chinese dialects

Almost all the Chinese dialects are developed from the same root and they have inherited many common features grammatically, lexically, phonologically and phonetically [14, 16]. For example, most of them are sharing the same written systems, very similar sound systems, the same phonological structures, and etc. In this thesis, dialects are compared mainly based on their phonological units, so the phonological features of Chinese dialects are introduced here.

In fact, all the Chinese dialects have the same phonological structure. Every written character is pronounced as a tonal mono-syllable and which can be divided into two kinds of phoneme sets, initials and finals, according to traditional Chinese phonology [29]. The initial is optional and always consists of an

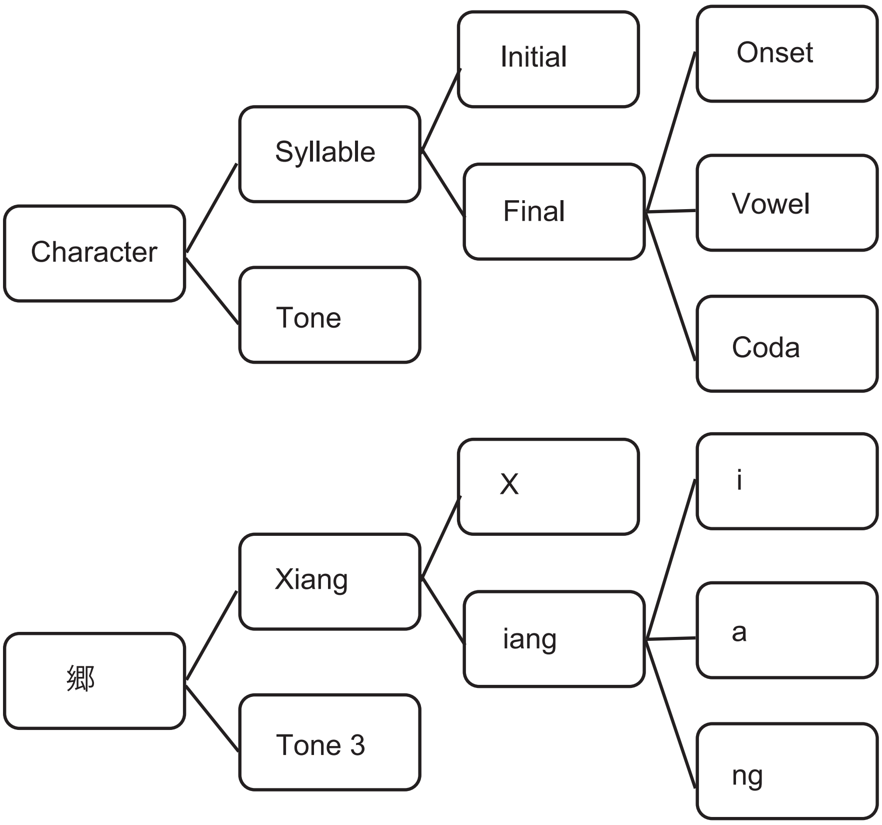


Figure 2.1: Phonological structure of Chinese

consonant. The final mainly consists of a vowel (which can be monophthong, diphthong or triphthong) with an optional onset or coda consonant. The tone is always carried by the final. The phonological structure of Chinese dialects and an example are shown in Fig. 2.1. By this figure, we can find that every character is pronounced as a monosyllable with tone, the syllable is combined an optional initial and a final, the final is combined by a vowel and optional onset and coda.

2.2.3 Different features among Chinese dialects

Due to many different historical, geographical reasons, there are many differences among Chinese dialects, with respect to grammar, vocabulary, and syntax. Therefore, the dialects of different major dialect regions are always mutually unintelligible to each other.

Take the phonological differences among Chinese dialects as example, their phonological inventories are quite different to each other [19]. In Mandarin, there are 22 initials (including null initial) and 38 finals. There are 19 initials and 53 finals in Cantonese, 35 initials and 32 finals in Wu dialect. Meanwhile, the phonemes of different dialects are also different. The codas of Mandarin are restricted to /n/ and /ŋ/, but some dialects have much more codas such as /m/, /p/, /t/, /k/. There are 5 tones in Mandarin including the neutral tone, but there are 9 tones in Cantonese and 2 tones in some dialects of Wu dialect region.

2.3 Current situation of Chinese spoken language

2.3.1 Popularization of standard Mandarin

Because people from different dialect regions always have difficulty in oral communication, standard Mandarin, which is the largest spoken language mainly spoken in northern and southwestern China, is taken as a common language of communication officially. Then standard Mandarin is popularized all over the country as the language used in government agencies, in the media and as the instruction language in schools. Therefore, most of native dialect speakers began to learn Mandarin. For the speakers from dialect regions where the dialects are mutual unintelligible to standard Mandarin, they are learning Mandarin just like a second language.

2.3.2 Development of accented Mandarin

Since the popularization of standard Mandarin, many dialect speakers are learning Mandarin just like learning a second language. But affected by their native dialects, many of them have some troubles in pronouncing some phonemes of standard Mandarin. For example, speakers from northeastern and southern China often mix up some phonemes like /zh/ and /z/, /ch/ and /c/, because their dialects don't make these distinctions. Therefore, many dialect speakers speak Mandarin with regional accents [30]. For some speakers from Xiang and Gan dialect regions, as the speakers cannot read some words in Mandarin, they just read these words in their dialects. Then some new local spoken languages are developed and which are in fact the mixture of the native dialects and standard Mandarin.

2.3.3 New developing trend

On the other hand, affected by the popularization of standard Mandarin, many dialects are losing some of their special features and becoming more and more similar to standard Mandarin. Further, with the development of China, many people are moving from one dialect region to another all over the country these years. When they move to a new dialect region, they will attempt to pick up the local dialects. However, affected by their native dialects, their pronunciations of the new dialects always show some features of their native dialects and strictly speaking, every of them have their own dialects.

Anyway, although Chinese dialects keep developing and losing some special features, it is still believed that they will continue to be used in the future, especially for the major dialects like Cantonese, Hakka and so on. Because even out their dialect regions, speakers will speak dialects to the people who are from the same dialect regions to show the special relationships between them and to get the strong group identity.

2.4 Conclusion

In this chapter, some fundamentals of Chinese dialects and the current situation of Chinese spoken languages are introduced. There are many different dialects, sub-dialects in China and they are mutual unintelligible to each other sometimes because they are different to varying degrees grammatically, lexically, phonologically and phonetically. Therefore, standard Mandarin has been popularized all over the country as official language and many dialect speakers begin to learn Mandarin just like a second language. However, affected by their native dialects, many of them speak Mandarin with regional accents. On the other hand, Chinese dialects are also developing affected by the popularization of Mandarin and some other dialects because people are moving from one dialect region to another all over the country. So in brief, the current situation of Chinese dialects are becoming more and more complicated and speakers from the same dialect regions may speak different dialects not only because they may from different sub-dialect regions but also because their dialects may already changed affected by Mandarin or other dialects.

Chapter 3

Related works

3.1 Introduction

In this chapter, background knowledge of modern speech processing and some related works of my study are introduced. First, the conventional framework of Automatic Speech Recognition (ASR) system is described and how to extract the acoustic features from speech is introduced. Then how multi-dialects are processed under this framework is discussed. After that, some related linguistic studies like capturing the phonetic differences using vowel normalization are introduced.

3.2 Modern speech processing of dialects

3.2.1 Conventional framework of ASR

The conventional framework of modern ASR system is shown by Fig 3.1. At the beginning, the input speech signal is preprocessed and segmented into a frame sequence by different windows. Then for every frame, acoustic features are extracted as feature vectors and passed to the decoder. After that, the feature vector sequences are treated as observations and decoded with the acoustic model and language model using algorithms like Viterbi or Baum-Welch.

Take the feature vector sequences as X, the corresponding word sequences as W, $P(W|X)$ as the posterior probability of observing vector sequences X when word sequence W is pronounced, then the decoder will try to get the maximum $P(W|X)$ and the corresponding $\hat{W}$. This process can be shown by

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X). \tag{3.1}$$

According to Bayes rule, the posterior probability can be calculated as the product of the class conditional distribution and the prior

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(X|W)P(W). \tag{3.2}$$

Here $P(X|W)$ is referred to as the score obtained by comparing the feature vector observations with the acoustic models, and $P(W)$ is referred to as the language score. Hence, the decoder incorporates acoustic models and language models to produce a word sequence that maximizes the posterior probability of the feature sequence. At last, the output optimal word sequence, the recognized result will be passed to the application.

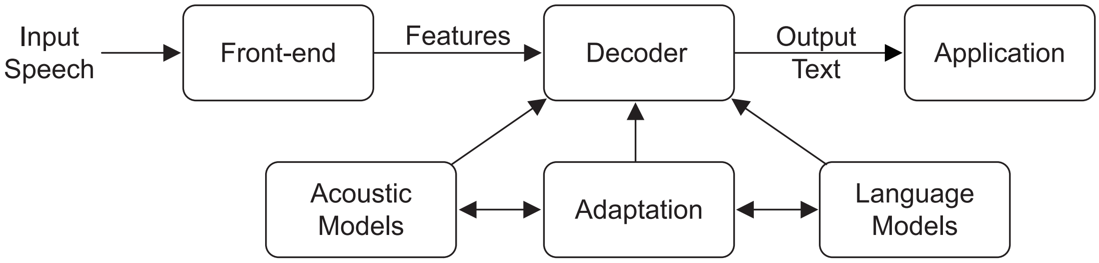


Figure 3.1: The basic framework of ASR

3.2.2 Acoustic features extraction

In the feature extraction stage, after the input speech signal is segmented into a sequence of frames, the acoustic features of them can be extracted using different methods.

Among the acoustic feature extracting methods, cepstrum is always used and this method is motivated by both perceptual and performance aspects of human

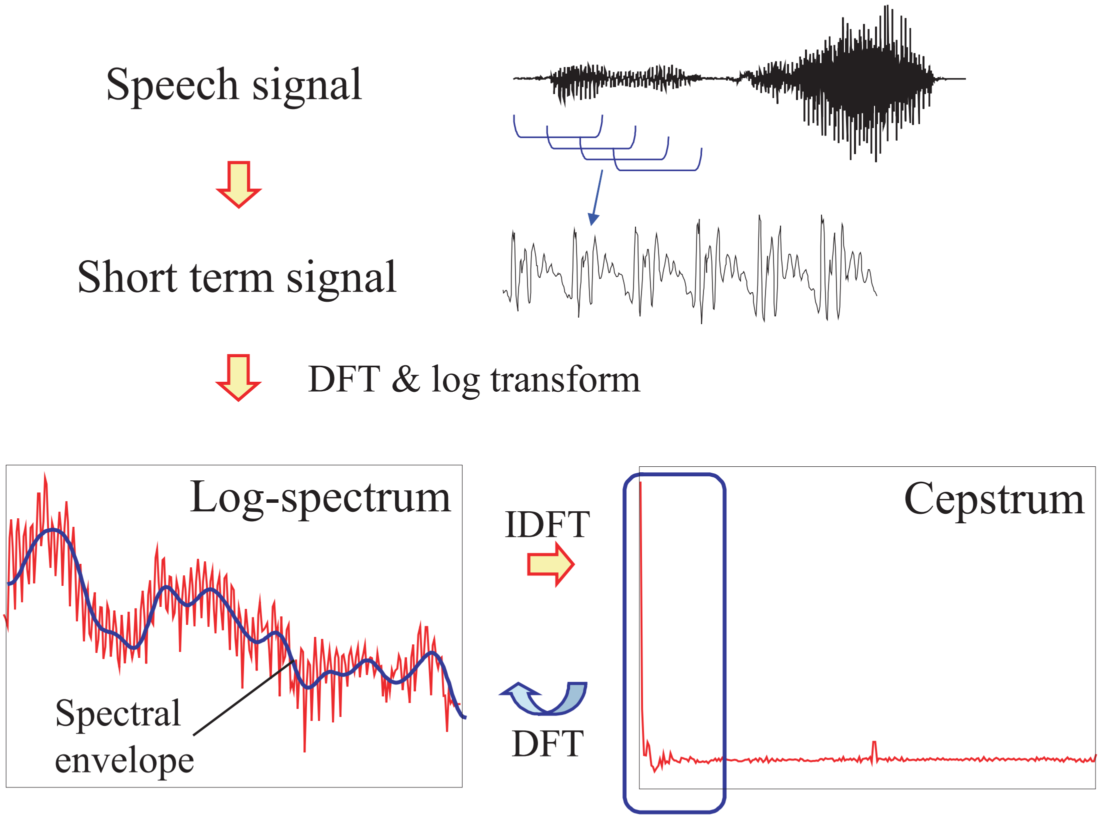


Figure 3.2: Cepstrum extraction from speech

beings. In speech production, the vocal tract may be viewed as a filter acting on a sound source [31, 32] and it changes shaper slowly in continuous speech. Therefore, at small enough time scales, the time scale of one frame can be considered a filter of fixed characteristics [33]. Hence, the process of extracting cepstrum from speech is shown by Fig 3.2. At the beginning, a short time Fourier transform is applied to convert the time domain signal into the frequency or spectral domain. Then a first-order pre-emphasis filter is usually applied to accentuate the higher frequencies in the formant structure. So the speech signal is windowed at intervals to produce discrete frames and Discrete Fourier Transform (DFT) is applied to compute the spectrum. After that, cepstrum can be obtained using the inverse DFT of the logarithm of the power spectrum. Then, spectral envelope can be extracted after DFT with a low pass filter. Its spectral peaks are defined as formants, which can be taken as the resonances of the human vocal tract. Therefore, the frequencies of formants are always used to represent the acoustic features and are usually used in some linguistic studies.

Nowadays, a new representation of the acoustic features of speech, Mel-Frequency Cepstrum Coefficient (MFCC) [37], is proposed and becomes the most commonly used representation of acoustic features. According to the perceptual aspects of listeners, Mel scale is proposed as the perceptual scales [38] and it warps the frequency scale by logarithmically compressing it. Therefore, different to cepstrum, MFCC is calculated using a series of triangular band-pass filters which are equally spaced on the Mel scales. The triangular band-pass filters are shown by Fig 3.3. Generally speaking, MFCC approximates the human auditory system's response more closely by integrating its spectral components over gradually widening intervals following the Mel scale and projecting the resulting Mel-warped spectrum on the cosine basis [36].

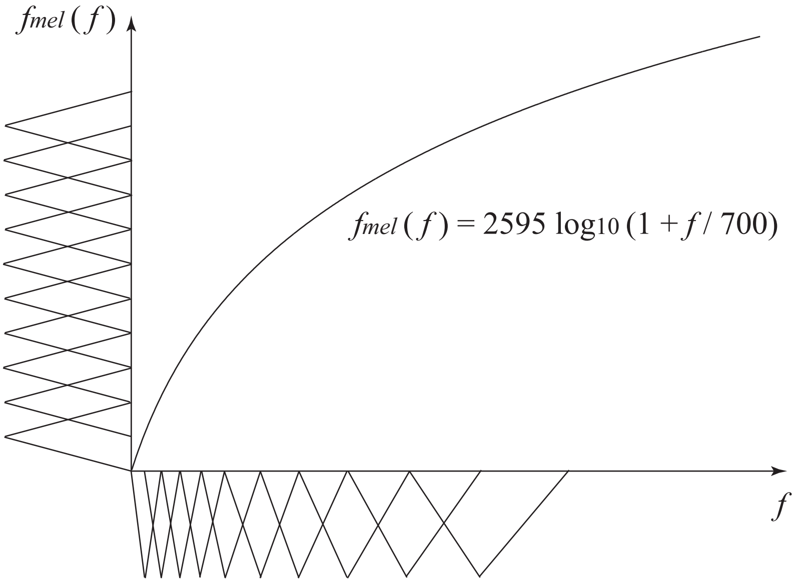


Figure 3.3: Triangular band-pass filters in calculating the MFCC

3.2.3 Acoustic modeling

After the acoustic features of speech are extracted as feature vectors, they will be compared with trained acoustic models to get the acoustic score $P(X|W)$. Nowadays, acoustic model is usually trained as Hidden Markov Model (HMM) [39] and one example of HMM is given by Fig 3.4. In this figure, S_i means the i_{th} state, a_i means the transition probability from state S_i to S_j, $b_i(x)$ is the output probability of state S_i generates the feature vector x. These parameters of HMM are learned in a data-driven manner and Maximum Likelihood (ML) criterion is always used. The optimal set of parameters θ should maximize the likelihood of the training data for the reference transcription W.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(W|\theta). \tag{3.3}$$

Generally speaking, Baum-Welch algorithm [40] is adopted for maximum-likelihood training of HMMs parameters using algorithm like EM [41]. By different iterations of increasing the likelihood of the training data, a local maximum is got and the HMMs parameters are fixed.

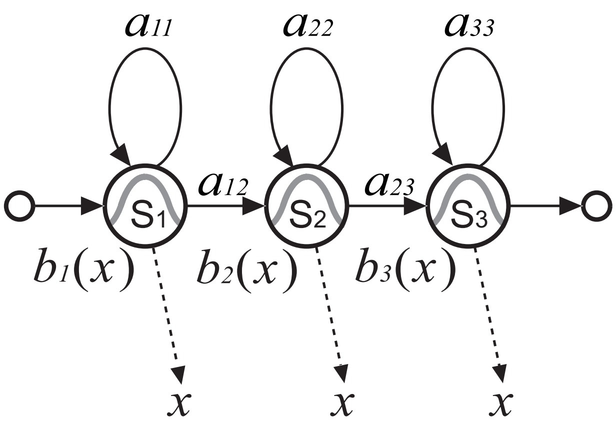


Figure 3.4: Hidden Markov Model (HMM)

In fact, for different recognition tasks, different kinds of HMMs can be built. Take the recognition of isolated word like digit recognition as example, HMM can be built for every word because we needn't to capture the co-articulatory effects between words and we can get sufficient training data to estimate the models. Then for the recognition system of large vocabulary, HMM will be trained for every phone, but heir acoustic features will be different depending on their neighboring phones. Therefore, content dependent models, HMMs based on the left and right contexts are usually adopted. Comparing the phone HMM without considering the context is called monophone, the phone HMM considering its right and left context is called triphone. Sometimes, more contexts, with two phones to the left and two to the right, will be considered and their name is quinphones [42].

3.2.4 Dialect processing

The process of one dialect is just like the common speech processing of one language. Theoretically speaking, after the acoustic and language models are trained with enough dialectal data, satisfactory performances can be got using some vocal tract length normalization or speaker adaptive techniques together. However, about the processing of multi-dialects or accented Mandarin [43], it is quite different and much more difficult.

The processing of multi-dialects can be divided into several different categories, like dialect identification, dialect speech recognition and so on. Meanwhile, these studies can be classified into two parts according to which kind of features, acoustic features or linguistic features, are mainly focused. For example, about some studies of language identification or dialect identification [44, 45, 46], the researchers are focusing on the linguistic features of the dialects and the dialect identification is achieved by building different language models. About the studies based on acoustic features, different kinds of phone recognizer can be built. Take [7] as example, parallel phone recognizers or adaptive phone recognizer can be built to recognize the dialect or dialect speech.

Generally speaking, if there are enough data for all the dialects, dialect-dependent but speaker-independent parallel recognizers are always built for multi-dialect recognition [47]. At the beginning, the dialect of input speech will be identified and be passed to the recognizer of this dialect. Then the results will be given. Otherwise, the input speech can be recognized by the recognizers in parallel and the input dialect will be determined by the recognizer that returns the highest ASR score. The process of these two methods can be shown by Fig 3.5.

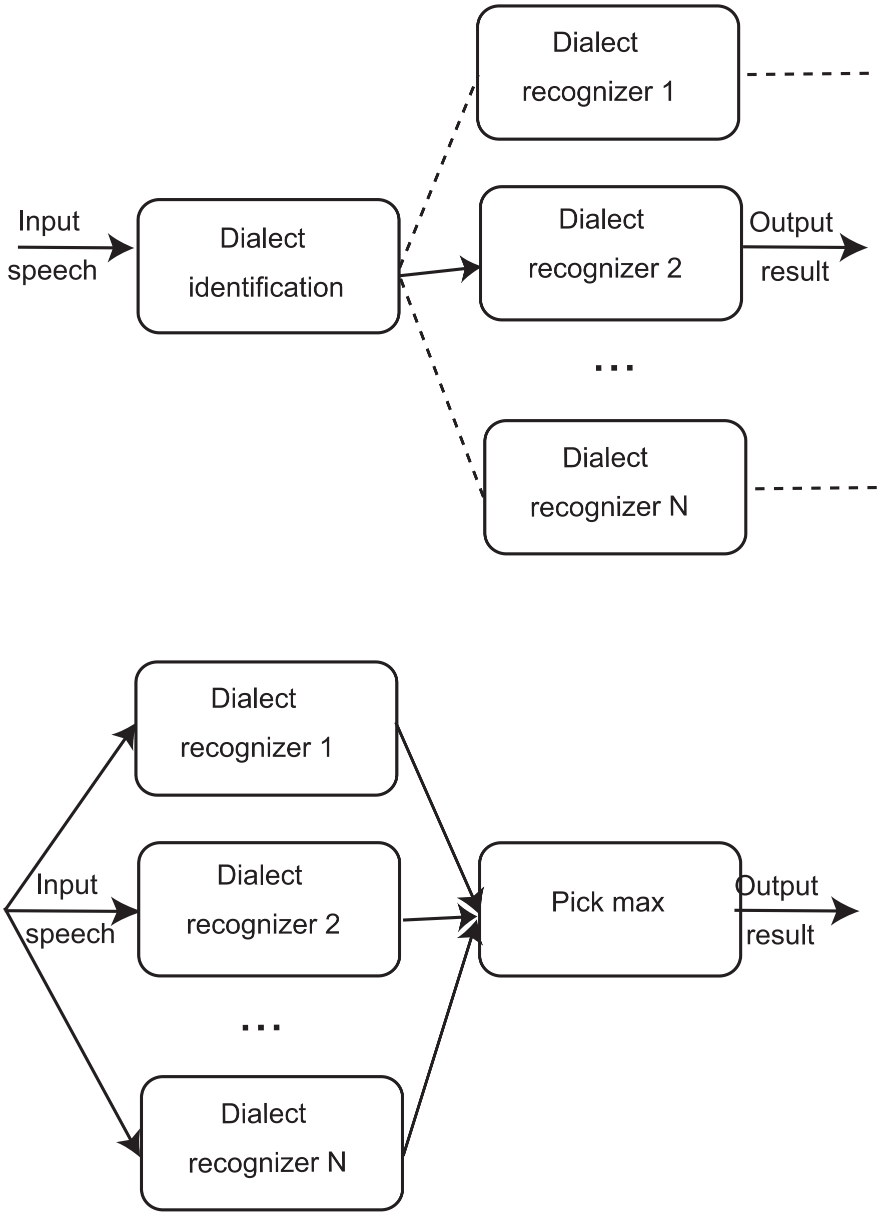


Figure 3.5: Two frameworks of multi-dialect recognition with multi-recognizers

3.2.5 Dealing with the extra-linguistic features

However, no matter which kind of acoustic feature representation is adopted, they are still affected by the extra-linguistic features such as speakers feature, recording microphone and so on. Therefore, normalization or adaptation techniques are usually used to deal with the different extra-linguistic features such as different speaker and recording conditions [48]. For example, in order to normalize the vocal tract length differences between male and female, vocal tract length normalization (VTLN) is always used [49]. And for more general speaker adaptation, maximum likelihood linear regression (MLLR) [50] is always used. Sometimes, in order to capture the speech dynamics, delta and delta-delta coefficients will be used together with techniques like heteroscedastic linear discriminant analysis (HLDA) [51]. Many systems also use some purely statistical approach to estimate the HMM parameters, like maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE) [51].

3.3 Related linguistic studies of dialects

3.3.1 New linguistic study of dialects

With the development of modern speech processing techniques, some new studies of dialects are started. For example, the acoustic features of dialect utterances are studied together with the articulatory features to find the relationships among these dialects. Generally speaking, the following acoustic features are always focused: the first several frequencies of the formants, the amplitude, the fundamental frequency and so on. The first several frequencies of the formants, generally the first three formants, are mainly characterized by the vocal tract shape and they are responsible for the major part of the information in speech. The acoustic studies of dialects are mainly based on these acoustic features. The amplitude and the fundamental frequency can be roughly taken as the pitch and the loudness of speech. They also account for suprasegmental information like stress and intonation. Sometimes, the length of phonemes and speed of the speech are also studies.

However, because different speakers have different vocal tract shapes and mouth sizes and these features characterize the formants of speech, the frequencies of the formants of the same vowel are always different to speakers. Especially for the speakers of different gender, the frequencies of their formants are quite different to each other. Fig 3.6 shows the frequencies of the first formant and the second formant of the 5 Japanese vowels. And about this figure, we can find that the formant frequencies of male and female speakers are quite different. Therefore, vowel normalization is needed to compare the vowel realizations by different speakers in meaningful linguistic and sociolinguistic ways. Furthermore, for the study of comparing the vowel realization of different dialects, another target of vowel normalization techniques should be achieved that it should tell how much of the differences is affected by the dialectal features and how much is affected by the physiological features of the speakers.

3.3.2 Vowel normalization techniques

When the phonetically equivalent vowels are produced by speakers of different genders, the formant frequencies are quite different. Generally speaking, the formant frequencies of adult female speakers are higher than those of adult male speakers largely because the vocal tract of female speakers are much shorter than male speakers. Therefore, vowel normalization is adopted to cancel such extra-linguistic features caused by physiological differences among speakers.

In [52], several general goals of utilizing vowel normalization techniques to eliminate the physiological differences among speakers are introduced: to preserve the purely linguistic differences in vowel quality; to preserve the phonological distinctions among vowels; to model the cognitive processes of human listeners. For sociolinguists and dialectologists, the last goal of modeling the cognitive processes of human listeners is the least important and preserving the phonological distinctions among vowels is not very important either. Their main target is to filter the extra-linguistic features caused by physiological differences to preserve sociolinguistic, dialectal or cross-linguistic differences in vowel quality. It is very similar to my study of classifying speakers based on the purely linguistic features of their dialect pronunciations by removing the extra-linguistic features.

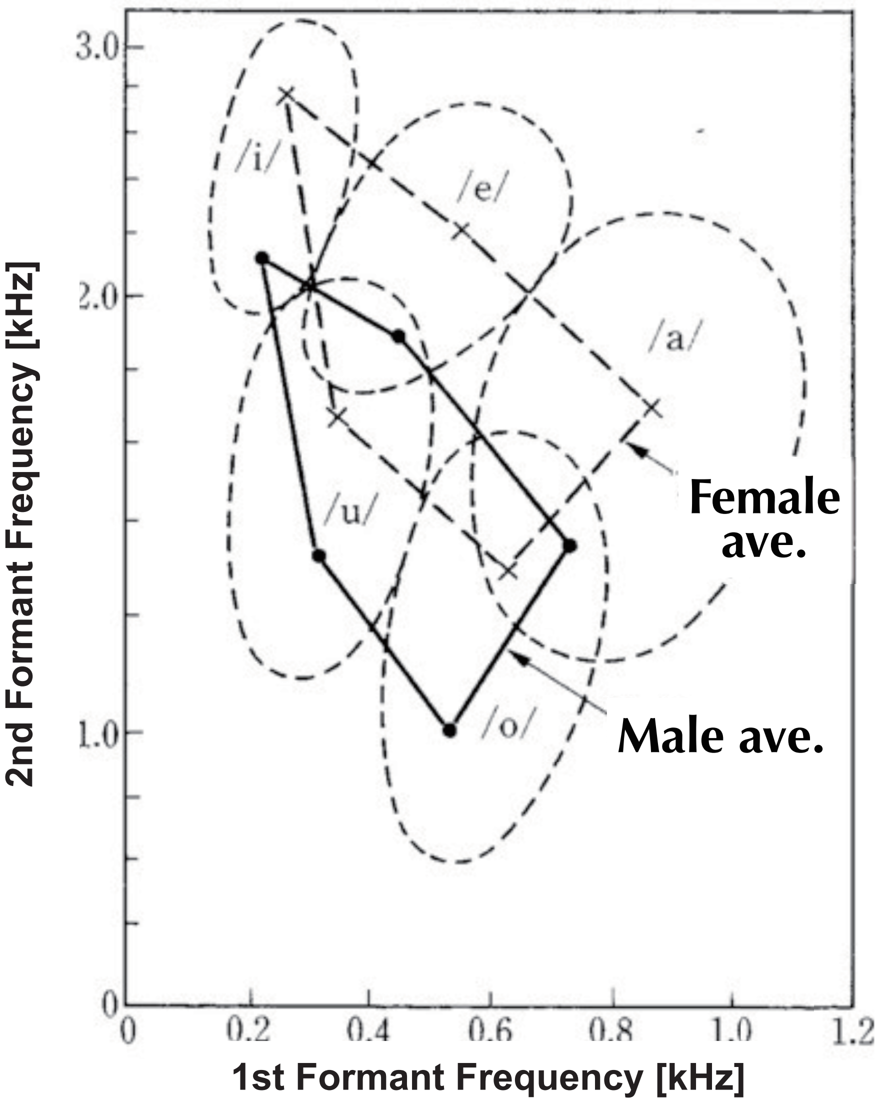


Figure 3.6: F1-F2 charts of Japanese vowels

According to the detailed techniques of vowel normalization, they fall into two general groups: vowel-intrinsic and vowel extrinsic. The vowel-intrinsic method always compare individual vowels using various combinations of formant values (F1, F2, usually F3, and occasionally F4), F0 (the fundamental frequency), or even formant bandwidths. [53] provided some useful history on vowel-intrinsic methods. Vowel-extrinsic methods, on the other hand, compare formant values of different vowels. [54] evaluated how well different formulas matched impressionistic transcriptions and came down solidly in favor of vowel-extrinsic methods, especially those by [55] and [56].

Meanwhile, the vowel normalization techniques can also be classified into two groups: speaker-intrinsic and speaker-extrinsic. Speaker-intrinsic methods normalize the vowels of a single speaker and speaker-extrinsic methods normalize the vowels of different speakers. In speaker-extrinsic techniques, some known properties such as the speaker's average formant values across a large number of vowels are always utilized for the normalization. For example, about the speaker-extrinsic methods used by [2], the log mean normalization technique developed by [56] is adopted and 134,000 vowels of 439 speakers are measured to derive the uniform scaling factor for the log mean of all formants. Therefore, on the other hand, this speaker-extrinsic method like that can not take account of any speaker specific properties.

3.3.3 Some other related studies

Nowadays, there are also some researchers focusing on Chinese dialects. For example, in the study of [57], they claim that although the formant frequencies of the same vowel are different to speakers, the relative positions of the vowels in the acoustic vowel chart are steady. So they calculate the relative values of the formant frequency and draw a vowel chart which is robust to the speaker features. After the values of F1 and F2 are converted into Bark, B1 and B2 can be obtained. Then using the new value, the following formulas are adopted to calculate the relative values of F1 and F2,

$$V1 = [(B1x - B1min)/(B1max - B1min)] \times 100 \tag{3.4}$$

$$V2 = [(B2x - B2min)/(B2max - B2min)] \times 100 \tag{3.5}$$

Where the $B1x$, $B2x$ means the current value, $B1max$ and $B1min$ means the maximum and minimum value of $B1x$. Then using the new $V1$ and $V2$, a relative vowel chart can be drawn. After that, the vowels of different dialects can be compared through the relative vowel chart. Meantime, as Chinese is a tonal language, they also proposed a method to calculate the relative F0 information using

$$T = [(x - min)/(max - min)] \times 5. \tag{3.6}$$

And this result is very similar with the traditional 5 level theory of Chinese tones.

There are also some researchers focusing on the relationships among Chinese dialects. By studying the traditional rhyme books like Qieyun (切韻) and Guangyun (広韻), which are wrote during the period of time from 6th to 10th century, some clues are found about the development of current Chinese dialects, especially the phonological features. Meantime, for many dialect cities, the native dialect data covering all the phonological inventories of that dialect are recorded. Then by comparing the native dialectal utterances of the different phonological units like initials, finals and tones, the current phonological difference among dialects are obtained. At last, according the results of the above studies, some handbooks about Chinese dialects are published [58, 59] and some written characters covering most of the phonological differences among dialects are fixed to be used as the reading materials for checking the pronunciation of different dialect speakers.

In the study of [60], considering the relations between articulatory features and acoustic features in vowel realizations and the fact that the articulatory features are not affected by physiological features of speakers, articulatory features are used to calculate the distances between any pair of corresponding phonological units and the phonological similarity between two Chinese dialects is calculated. At the beginning, for every kind of phonological units, the differences between two phonemes are classified into several different levels according to some fixed articulatory features. Take the initials as example, the following articulatory features are checked: the articulatory methods (stop, nasal, glide, liquid..), voiced or unvoiced, fricative or not, aspirate or not, the articulatory organs, whether the lips are used, whether they are retroflex consonants, the stop or affricative position and so on. About any two initials, the above articulatory features of them are checked and a contrary degree between them is calculated by finding how many kinds of the above articulatory features are different to each other. Then the contrary degree between two syllables can be obtained by summing the contrary degrees between their initials, finals and tones. At last, the similarities among the dialects are calculated by the total contrary degrees of the corresponding syllables.

3.4 Conclusion

In this chapter, some related works are introduced. First, the conventional framework of Automatic Speech Recognition (ASR) system and how to extract the acoustic features and build acoustic models are introduced. Then some related works about multi-dialects processing are introduced. After that, some related linguistic studies of eliminating extra-linguistic features caused by physiological difference among speakers using vowel normalization are introduced and some related linguistic analysis of Chinese dialects are presented at last.

About my study of classifying speakers based on their dialect pronunciations, the methods introduced above may not works in this problem. For example, if we try to apply the method of building dialect-dependent but speaker-independent acoustic models to dialect-based speaker classification, we must build several dozens of dialect or sub-dialect models and collect the data of many speakers from the same dialect regions. It will be a very challenging work and this approach will also conflict with the goal of finding the intra-dialect relations among speakers, because two Chinese speakers of the same dialect may be speakers from different sub-dialect regions. Therefore, in my study, a new structural pronunciation representation of Chinese dialects is proposed to extract the purely linguistic features to classify speakers based on their dialects.

Chapter 4

Structural representation of Chinese dialects

4.1 Introduction

In this chapter, how to build Chinese dialect pronunciation structures and how to calculate the distances between two structures are introduced. After the segmental features of speech are represented by spectrum, the extra-linguistic features in speech can be modeled as affine transformations. Then, for every speaker, by calculating the Bhattacharyya Distances (BDs) between any pair of his/her utterances, pronunciation structure can be built. Because BD is invariant to affine transformations, the built structure is invariant to extra-linguistic features in speech. After that, how to build comparable dialect pronunciation structures is also introduced. According to the studies of some linguists who are focusing on Chinese dialects, some lists of characters covering the dialect differences are fixed to check the dialect of different speakers. Then using these lists as reading materials, comparable dialect structures are built. At last, how to calculate the distances between two comparable structures is introduced.

4.2 Mathematical model of extra-linguistic features

In modern speech processing techniques, the segmental feature of speech is always represented by spectrum and it contains not only linguistic feature, but also extra-linguistic features corresponding to age, gender, microphone, recording background and so on. However, for the problem of dialect-based speaker classification, only linguistic features are needed and the extra-linguistic features should be canceled or removed.

In fact, after speech is represented by spectrum, the extra-linguistic features can be classified into three kinds according to the distortions they cause in spectral space: additive distortion, convolutional distortion and linear distortion. The additive distortion is always caused by extra-linguistic features such as background noises. As this kind of features can be avoided by changing the recording environment, we just focus the next two kinds extra-linguistic features. The convolutional distortions are always caused by extra-linguistic features like recording microphone and transmission lines. Sometimes, the vocal tract shapers of speakers also cause a convolutional distortion in spectral space. The vocal tract length is a typical reason of linear transformational distortion, which is already shown by [61].

If a speech event is represented by a cepstrum vector c, the convolutional distortion can be represented as addition of another vector b and changes c into $c' = c + b$. Meanwhile, the linear transformational distortion is modeled as frequency warping of the log spectrum and changes c into $c' = Ac$. So the total distortions caused by inevitable extra-linguistic features can be modeled by $c' = Ac + b$, known as affine transformation. These distortions can be schematized by Fig. 4.1, where the horizontal and vertical distortions correspond to the distortions due to matrix A and vector b, respectively.

4.3 Speaker-invariant dialect structures

The spectral distortions caused by extra-linguistic features can be modeled as affine transformations. So if we can build an acoustic structure which is invariant

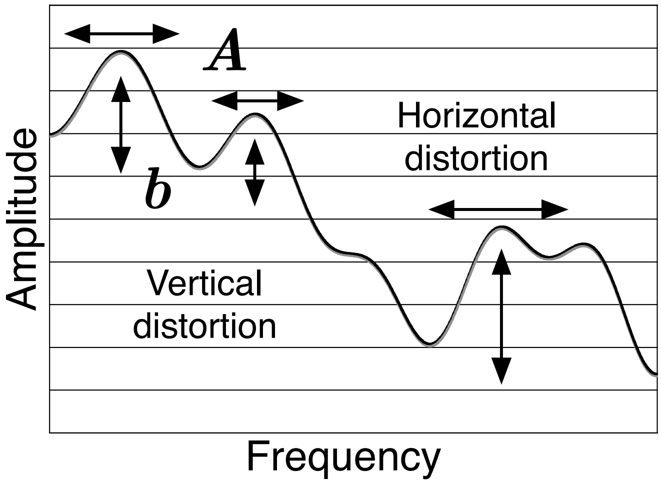


Figure 4.1: Spectral distortions caused by matix A and vector b

to affine transformations, it will be invariant to extra-linguistic features. In [24], it is proved that Bhattacharyya Distance (BD) is invariant to affine transformations. Then, we can use BD to build pronunciation structure and it is invariant to extra-linguistic features.

Here, every speech event is captured as a distribution ($p_i(c)$) and event-to-event distances are calculated as Bhattacharyya Distance (BD).

$$BD(p_i(c), p_j(c)) = -\ln \oint \sqrt{p_i(c)p_j(c)}dc, \tag{4.1}$$

By calculating BDs between any pair of speech events, a distance matrix can be obtained. Since a distance matrix can represent uniquely a geometrical shape composed of all the speech events, we call the matrix a pronunciation structure of these speech events. And because BD is invariant with respect to affine transformations, the obtained pronunciation structure is invariant to extra-linguistic features.

Fig. 4.2 shows an example of the invariant underlying structure among three sets of speech events. Any set of the events are obtained by affine transform of either of the other two sets. This means that the BD-based distance matrix is invariant and common among the three sets. So if the pronunciation structures are built separately from two speakers of the same dialect, structural difference between them is small. If they are built from a single speaker who can speak different dialects, the difference will be large but it is independent of age and gender.

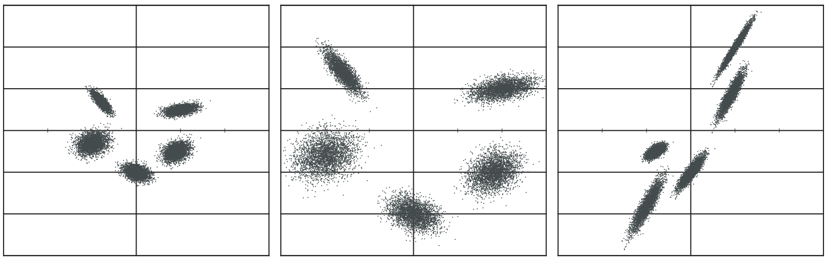


Figure 4.2: The invariant underlying structure among three data sets

4.4 Comparable dialect pronunciation structures

In order to analyze the pronunciation of speakers from different dialects using structural representation, comparable dialect structures have to be built for these speakers using their dialectal utterances of the same set of linguistic units. Considering that there are many grammatical and lexical differences among Chinese dialects, syllable or smaller phonological units can be a good choice. However, although all Chinese dialects are sharing the same phonological structures, the inventories of their phonological units change from dialect to dialect, so phoneme-based pronunciation structures cannot be built. Then considering that all the Chinese dialects are sharing the same written characters and every character is pronounced as a mono-syllable, the utterances of syllable units (characters) become the best choice to build the pronunciation structure for dialect comparison.

After the characters become the best choice of reading materials to build the comparable dialect pronunciation structures, the problem becomes how to select the characters. In fact, for different purposes, different characters can be selected to build different comparable pronunciation structures. For example, in order to analyze the pronunciation of different dialect speakers, the characters covering the dialectal differences can be adopted, and in order to assess the accented Mandarin pronunciation of dialect speakers, the characters which are easily mispronounced can be adopted. Then with the pronunciation structures built using these data, the pronunciation of these speakers can be analyzed and assessed purely on their linguistic features which are invariant with extra-linguistic factors.

Recently, some dialectologists are focusing on the relationships among Chinese dialects by their phonological features. For example, using the dialectal utterances of the same written characters, the initial/final units of different dialects are listed and their phonetic features are compared. As a result, the relations of these units between Mandarin and other dialects are often shown. In [19], all the initial/final units in different dialects and their corresponding ones in Mandarin are listed together with some characters as examples. And in [58, 59], some specific lists of written characters are proposed to check the phonological differences among dialects. For example, in the latter one, three different lists of written characters are fixed for checking the dialectal features of tones, initials and finals, separately. Therefore, using these lists, different comparable dialect pronunciation structures can be built to check these different features among dialects.

4.5 Distances between pronunciation structures

After the dialect pronunciation structures are built for the speakers, the distance between the dialects of any two speakers can be calculated as the distance between their pronunciation structures. In [62], it is shown theoretically and experimentally that the vocal tract length differences rotates the pronunciation structure. So here, the distance between two structures is obtained after one is shifted ($+b$) and rotated ($\times A$) until the best overlap is observed between them, which is shown in Fig. 4.3. Then with the best overlap after shift and rotation, the distance between two structures is calculated as the minimum sum of the distances between the corresponding two events of the two structures. In [21], it was experimentally proved that the minimum sum can be approximately calculated as Euclidean distance between two distance matrices. Following is the detailed computing formula:

$$D_1(S,T) = \sqrt{\frac{1}{M}\sum_{i<j}(S_{ij} - T_{ij})^2}, \tag{4.2}$$

where S_{ij} and T_{ij} mean the (i, j) element of matrices S and T, respectively. M means the number of the speech events.

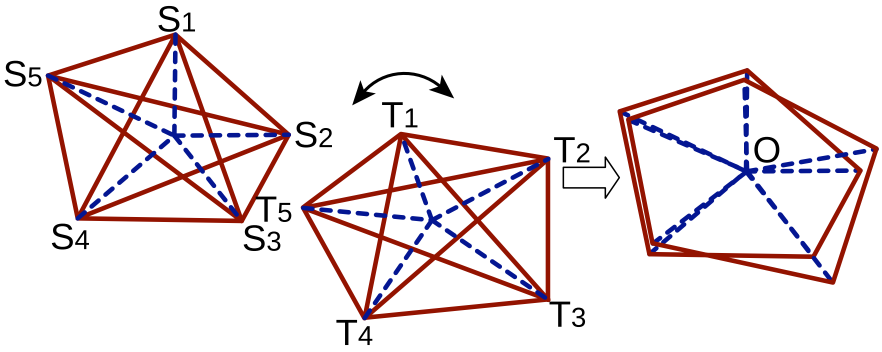


Figure 4.3: Distance calculation after shift and rotation

4.6 Conclusion

In this chapter, the extra-linguistic features contained in speech are classified by their spectral behaviors and the distortions caused by inevitable extra-linguistic features are modeled as affine transformations mathematically. Then using the affine-invariant feature of BD, a dialect pronunciation structure is built for every speaker by calculating the BDs between any pair of his/her utterances and it is invariant to extra-linguistic features. After that, I introduced how to select the proper reading materials to build comparable dialect pronunciations and how to calculate the distance between two structures.

Chapter 5

Speaker classification based on dialects

5.1 Introduction

In this chapter, a dialect-based speaker classification experiment using structural method is carried out and the result shows that Chinese dialect speakers can be classified based on their dialects using the structural representation of dialect pronunciations to extract the purely dialectal features. At the beginning, as publicly available Chinese dialect corpora cover only two or three dialects and cannot be used for my purpose, some recordings are carried out and the data of 18 Chinese dialect speakers are recorded. Then every syllable is taken as a speech event and calculated as a distribution. After that, for every speaker, Bhattacharyya Distance (BD) between any two speech events is calculated and the dialect pronunciation structure is built. Then these speakers are classified based on their dialects by calculating the distances between their dialect pronunciation structures. At last, the result shows that the speakers are well classified by their dialects and it is robust to speaker features.

5.2 Experimental data of dialects

5.2.1 Chinese dialect corpora

Nowadays, many researchers are focusing on the study of Chinese dialects and some dialect corpora are built. However, most of their studies are about only one or two dialects such as Cantonese, Shanghainese and these corpora cannot be used in our problem of dialect-based speaker classification. About the corpora covering different dialects, they are mainly built for the studies of some linguists and the reading materials, the recording environments are always different. Therefore, these corpora cannot be used in speech processing like our problem either and we have to record some new data of multi-dialects using the same reading materials.

Table 5.1: Detailed information of the dialect speakers

Speaker ID	Dialect	Hometown	Gender
01	Kejia	DaBo	M
02	Kejia	ShenZhen	F
03	Yue	FoShan	M
04	Yue	MeiXian	F
05	Yue	HongKong	M
06	Yue	HongKong	F
07	Yue	ShenZhen	F
08	Min	ZhangZhou	M
09	Min	FuZhou	F
10	Min	JiJiang	M
11	Wu	ShangHai	M
12	Wu	ShangHai	M
13	Wu	ShangHai	M
14	Wu	ShangHai	F
15	Wu	ShaoXing	M
16	Wu	NingBo	M
17	Wu	YiXing	M
18	Wu	SuZhou	F

5.2.2 Recording subjects

For this experiment, the recordings were carried out in Japan and 18 Chinese dialect speakers participated in the recordings. They were all graduate students in the University of Tokyo. The language backgrounds of these speakers were all checked to ensure that they and their parents were all born and brought up in the same dialect regions except one female speaker. Her parents are both native Hakka speakers and they moved to a Cantonese region when she was 10 years old. Therefore, she has mastered two dialects, Hakka and Cantonese, and her utterances of these two dialects have different linguistic features but the same speaker feature. In Table 5.1, more details like the hometowns and the genders of the speakers are listed together with their corresponding speaker IDs. The colors mean different dialect regions. The above mentioned female speaker has two speaker IDs, 02 and 07, which stand for her Hakka and Cantonese, respectively.

5.2.3 Reading materials and recording

For this experiment, 38 characters, which are covering all the 38 finals in Mandarin, were selected as the reading materials. These characters and their corresponding syllables of Mandarin are listed in Table 5.2.

Table 5.2: Selected characters and their pronunciations in Mandarin

Characters	辣、架、蛇、落、野、 月、資、知、耳、第、虚、 木、北、桂、桃、薬、流、 三、根、温、林、減、園、 隣、党、講、東、瓮、瓊
Mandarin Pronunciation	/bi/,/ci/,/shi/,/er/,/wu/,/yü/,/a/,/bo/,/e/, /ai/,/bei/,/zao/,/rou/,/zuo/,/ya/,/wa/,/bie/, /yue/,/uai/,/dui/,/yao/,/niu/,/an/,/yan/,/wan/, /juan/,/en/,/bin/,/wen/,/jun/,/ang/,/yang/, /wang/,/beng/,/bing/,/weng/,/zong/,/yong/

Table 5.3: Acoustic analysis condition

Sampling	16bit / 16kHz
Windows	Blackman, 25ms length, 1ms shift
Parameters	Mel-cepstrum, 1-10 Dimensions
Distribution	Diagonal Gaussian estimated with MAP

The recordings were carried out in a sound proof room, so the data are all expected to be clean. Before the recordings, the reading materials were checked by the speakers to ensure they can read them in their native dialects correctly. The recording equipments included a high quality microphone fixed on the table, a linear PCM recorder of Sony company. The data was recorded as monophony and the sample rate is 44 kHz. The speakers were asked to read the selected characters in their native dialects and each character was read four times. After that, every syllable was labeled manually according to their spectral performances. Every syllable is cut and stored into individual files. Then, these data were analyzed under the acoustic conditions shown in Table 5.3. Each speech event (final or syllable) was modeled as a diagonal Gaussian distribution and the parameter estimation was done for Gaussian modeling using Maximum A Posteriori (MAP) criterion.

5.3 Pronunciation structures shown by phonetic trees

In Mandarin, there are 9 monophthong finals and these finals are covered by the first 9 selected characters of Table 5.2. Using the recorded data, monophthong pronunciation structure can be built for every speaker after the final parts of his/her utterances of these 9 characters are cut and calculated as distributions. At the beginning, the utterances of these characters are labeled to finals. Then after every final part is cut and modeled as a single Gaussian individually, the BDs of every pair of monophthongs are calculated for each speaker and the monophthong pronunciation structure is built. At last, the monophthong structure, the distance matrix of the monophthongs, can be visualized as a tree diagram using Ward's clustering method [63].

Fig. 5.1 shows the phonetic trees of three Cantonese speakers of 03, 05 and 06. Speaker 03 is a male from FoShan, speaker 05 is a male from HongKong, and speaker 06 is a female from HongKong, too. In the figures, the nodes are the IPA symbols of the 9 monophthongs in Mandarin which represent the utterances of the monophthongs of every speaker. Then by the results, we can see that the phonetic trees of speaker 03 and 05 are structurally very similar but slightly different. Locally speaking, a difference between ɤ and uɔ in 03 is larger than in 05. Globally speaking, they are very similar considering that the mirrored position of the two sub-trees can be ignored in tree diagrams. Meanwhile, we can see the phonetic trees of speaker 05 and 06 are almost the same although they are different gender. It is because they are from the same city and their dialects are supposed to be the same. So the results show the phonetic trees of these speakers, the monophthong pronunciation structures, are sensitive to dialectal information and highly independent of genders.

5.4 Dialect-based classification experiment

In Fig. 5.1, the structures are obtained using the monophthong finals of the dialect syllable utterances. In fact, more dialectal features can be found by syllable-based analysis, where syllable-to-syllable distances have to be calculated. There are two methods to calculate this distance. One method is that a whole syllable is modeled as a Gaussian, just as in building the monophthong structures, each monophthong segment was modeled as Gaussian. Another is that a syllable is modeled as a sequence of a fixed number of distributions, such as HMM. Syllable-to-syllable distance is obtained as summation of distances between the corresponding distributions. Since these Chinese syllables are all very short, the first method is adopted to built the syllable pronunciation structures.

Using the syllable parts of the recorded data, every syllable is calculated as a distribution. Then for every speaker, dialect pronunciation structure is built by the BDs between any pair of distributions of that speaker. After that, the distance between two structures is calculated using the method in last chapter and the speakers are classified based the distances. At last, the classification result is shown by Fig. 5.2 using Ward's clustering method. In this figure, every speaker is represented by the speaker ID in Table 5.1 and the colors of the dialect regions are also the same.

In Fig. 5.2, we can find that the speakers from the same dialect regions are clustered together and the speakers from the same sub-dialect regions are also clustered near to each other. For example, speakers 11-14 from ShangHai are classified together and speakers 15-18 from different cities of Wu dialect region are classified near to speaker 11-14. About the other speakers, speakers from Min, Yue and Kejia dialects regions, are clustered in the other big tree. The result also shows high independence of extra-linguistic factors such as the genders of the speakers. For example, as described before, 02 and 07 represent different dialects of the same speaker and they are classified into their corresponding dialect groups correctly, not be classified into the same group. So, the result shows that using the pronunciation structures, the dialect speakers can be well classified by their dialects and which is highly independent of ages and genders of speakers.

5.5 Conclusion

Dialect-based speaker classification experiments are introduced in this chapter and the results show that the structural method can classify speakers based on their dialects by extracting the linguistic features. After the data of 18 native Chinese dialect speakers are recorded, their monophthong pronunciation structures are built and shown by phonetic trees. Then by the distances between their pronunciation structures, classification experiment of these speakers is carried out. All the results show that our structural method works well in dialect-based speaker classification and it is highly independent of extra-linguistic features like speaker features.

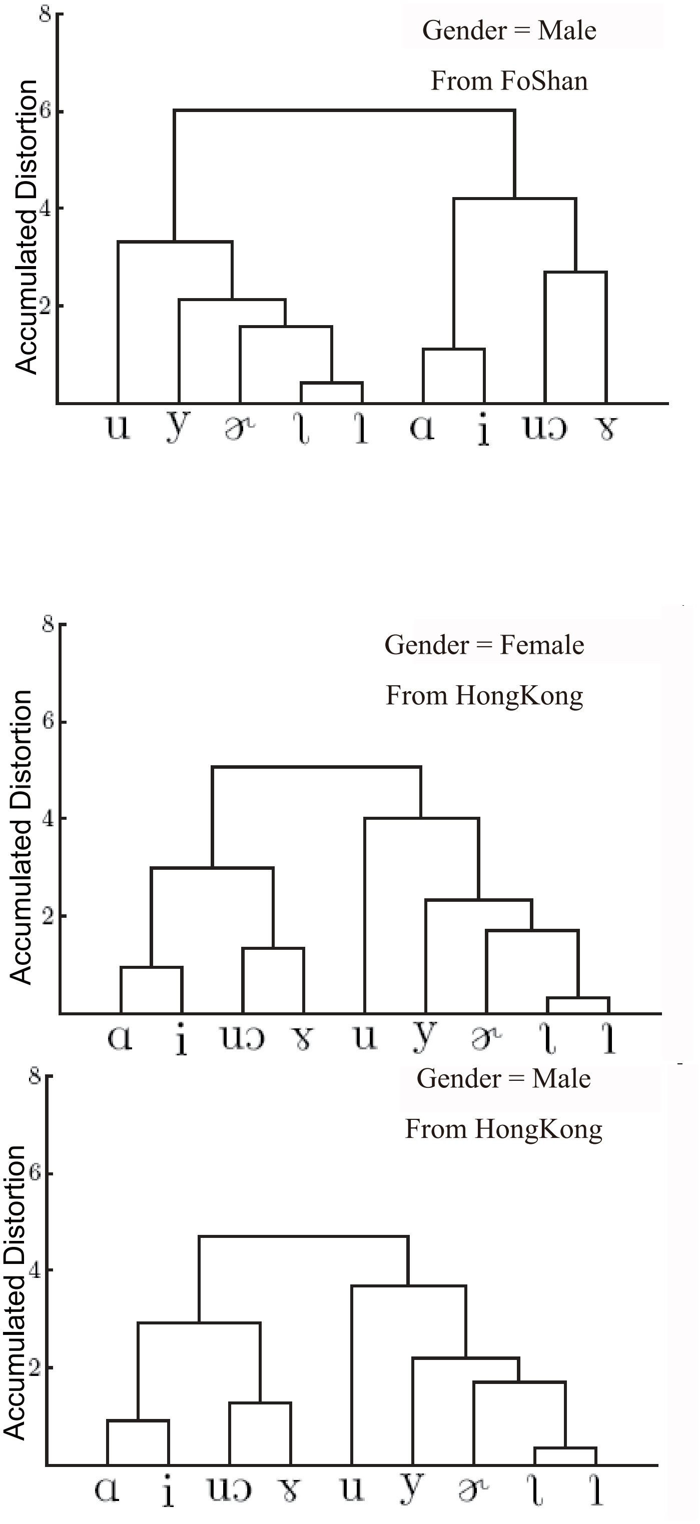


Figure 5.1: Phonetic trees of three Cantonese speakers

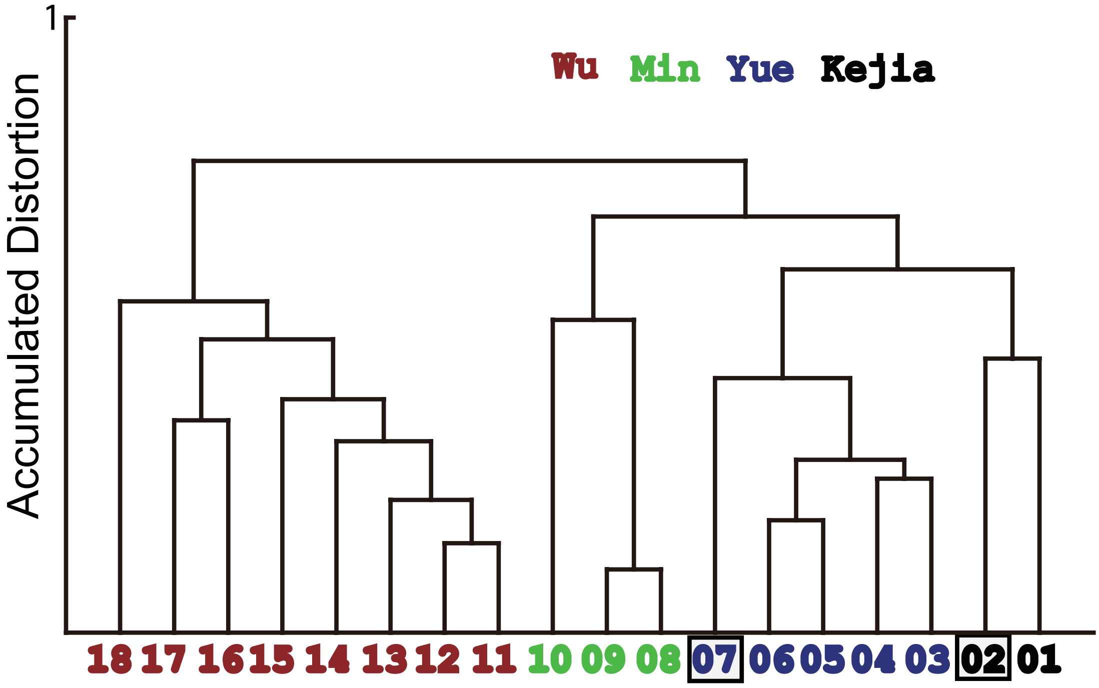


Figure 5.2: Classification of the dialect speakers

Chapter 6

Speaker classification based on sub-dialects

6.1 Introduction

In last chapter, it is shown that the structural representation of pronunciation can extract the purely linguistic features from Chinese dialects and classify speakers based on their dialects. However, that experiment was carried out using the data of speakers from several major dialect regions which are quite different to each other. Therefore, some comments may be given that this task of classifying speakers based on their major dialects is too easy.

In this chapter, the structural method is verified by a new sub-dialect-based speaker classification experiment. After the data of 16 speakers from 4 sub-dialect regions of Mandarin are recorded in China, sub-dialect based speaker classification experiment is carried out by building sub-dialect structures to extract the purely linguistic features. After that, the classification experiment is carried out and the speakers are classified based on the sub-sub-dialects of their hometowns. The results show that the speakers can be well classified based on their sub-dialect or sub-sub-dialects.

Table 6.1: Detailed information of the speakers

Speaker ID	Sub-Dialect	Hometown	Gender
01	XiNan	ChengDu	F
02	XiNan	ChengDu	F
03	XiNan	ChengDu	M
04	XiNan	ChengDu	F
05	JiLu	ShangQiu	F
06	JiLu	ShangQiu	F
07	JiLu	YuZhou	F
08	JiLu	YuZhou	F
09	BeiFang	TianJin	F
10	BeiFang	TianJin	M
11	BeiFang	TianJin	F
12	BeiFang	TianJin	M
13	JiaoLiao	YanTai	F
14	JiaoLiao	WeiHai	F
15	JiaoLiao	RuShan	F
16	JiaoLiao	RongChen	F

6.2 Experimental data of Mandarin sub-dialects

For the new experiment of classifying speakers based on their sub-dialects, I went back to China and recorded some data of Mandarin sub-dialects. The recording was carried out in Nankai University in Tianjin, China. The recording subjects are 16 speakers who are from 8 cities belonging to 4 sub-dialect regions of Mandarin. These speakers were selected after their language backgrounds were checked to ensure they were brought up in the same sub-dialect regions and their parents were also the native speakers of that sub-dialect. They are mainly graduate students in that university and have no background of other dialects or sub-dialects before entering the university. Here, every speaker was given a speaker ID and some other information of these speakers can be found in Table 6.1.

Table 6.2: Examples of selected characters

Characters	爬, 辣, 架, 夾, 花, 刮, 河, 色,..., 瓊, 胸
Syllables	/pa/, /la/, /jia/, /jia/, /hua/, /gua/, /he/, /se/, ..., /qiong/, /xiong/
Finals	/a/, /a/, /ia/, /ia/, /ua/, /ua/, /e/, /e/, ..., /iong/, /iong/

For this experiment, more than one hundred characters were adopted as the reading material in order to build the sub-dialect sensitive pronunciation structures. Some examples of these characters and their corresponding syllables and finals are listed in Table 6.2. The recordings were carried out in a quite room. The recording equipments included a high quality microphone fixed on the table, a linear PCM recorder of Sony company. The data was recorded as monophony and the sample rate is 44 kHz. During the recording, every speaker was asked to read the selected characters in their native sub-dialects three times.

After the recording, the data was labeled phonetically and manually by linguistic students. Then the final part of every syllable was modeled as a single Gaussian distribution under the acoustic conditions shown in Table 5.3 and the pronunciation structure for every speaker was calculated using the BDs between any pair corresponding finals.

6.3 Sub-dialect based speaker classification

After the sub-dialect pronunciation structure is built for every speaker, these speakers can be classified by calculating the distances between their pronunciation structures. The result is shown by Fig. 6.1, while the ID of every node is the same as that in Table 6.1 and the colors mean different sub-dialect regions.

In this figure, we can find that the speakers are mainly classified by their sub-dialects and the speakers from the same city are all classified together. Speakers 01-04, who are from XiNan sub-dialect region of Mandarin, are grouped together in a sub-tree. The speakers 09-12 and 13-16, who are from BeiFang and JiaoLiao sub-dialect regions, are clustered to two sub-trees respectively. But for the speakers from JiLu sub-dialect region, although speakers 05-06 from YuZhou and speakers 07-08 from ShangQiu are still grouped near to each other separately, they are finally clustered into different sub-trees. Speakers 05-06 are clustered near to the BeiFang sub-dialect region and speakers 07-08 are clustered near to the JiaoLiao sub-dialect region. In fact, these three big sub-dialect regions of Mandarin are not only very near to each other geographically, but also very near to each other linguistically [19]. And according to [19] and [20], the phonological differences among these sub-dialects regions of Mandarin are mainly based on the following three features: the tones, the pronunciation of alveolar initials (/n/, /l/, /z/, /c/, /s/), the pronunciation of retroflex initials (/zh/, /ch/, /sh/, /r/) and pronunciation of finals nasal with coda (/ng/, /n/). But in this experiment, only the finals are adopted and their pronunciation of the finals with nasal coda are generally the same in these three sub-dialect regions. Meanwhile, it is also considered as a reason that the sub-dialects of these speakers are affected more or less by other sub-dialects, because these speakers have been in the university, the sub-dialect region of BeiFang, for several years. Further, the sub-dialect regions of the speakers are obtained by traditional linguistics and the linguistic distances of dialects are different to the acoustic distances used in our experiment, which is also considered as a possible reason. Anyway, none of these reasons can be proved, so new comparison experiments are designed in the next chapter. About the result of this experiment, totally speaking, these speakers are mainly classified by their sub-dialects and speakers from the same cities are all classified near to each other.

6.4 Sub-sub-dialect based speaker classification

By last experiment, it is shows that speakers are mainly classified based on their sub-dialects by building the sub-dialect pronunciation structures to extract the dialectal features from their pronunciations. Then here, I will show wether the structure method can be applied to calculating the sub-dialect distances among hometowns of the speakers.

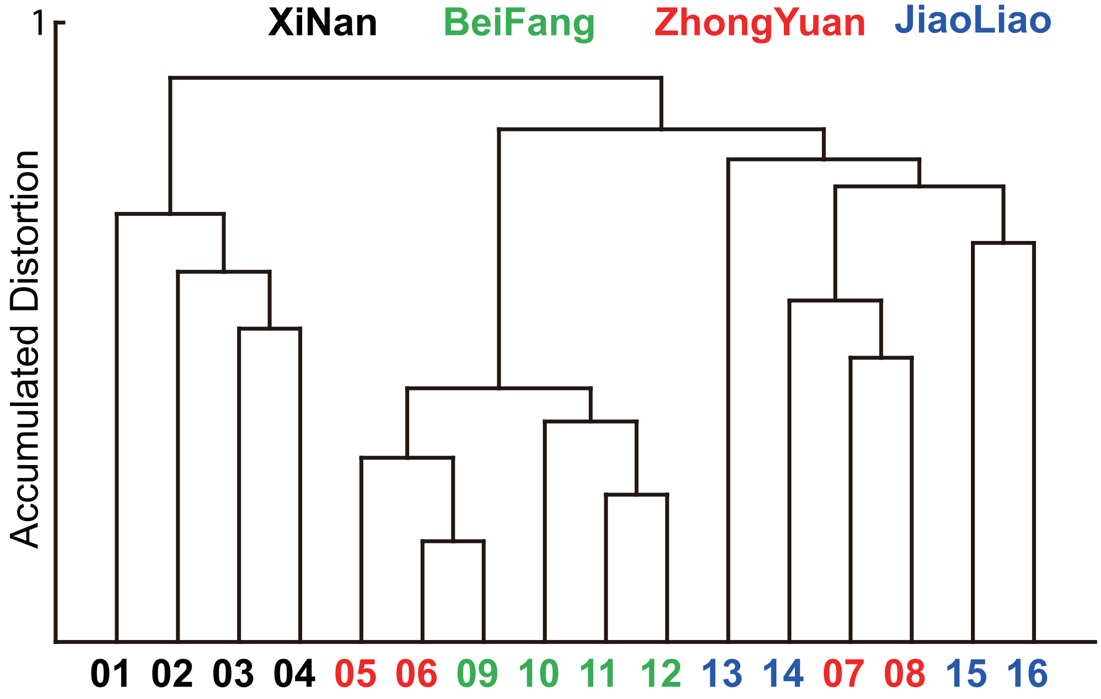


Figure 6.1: Sub-dialect-based speaker classification

Strictly speaking, although two cities belong to the same sub-dialect region, the pronunciations are somewhat different to each other. Here, the 8 hometowns of the 4 sub-dialects are considered to stand for 8 sub-sub-dialects. Then by building the pronunciation structure through averaging the structures of the speakers belonging to the same sub-sub-dialects (hometowns), the inter-town distances are calculated and hometown based classification is obtained. The result of this experiment is shown in Fig. 6.2, where every hometown is represented by the first two letters of their names. Referring to the dialect maps published by Chinese Academy of Social Sciences [28], further information on the sub-sub-dialects spoken in these cities can be obtained. The four cities (RC, RS, WH, YT) belong to the same sub-sub-dialect region, YZ and SQ belong to two different sub-sub-dialect regions, CD and TJ belong to different sub-dialects. Although we are only focusing on the acoustic features of finals, our results are similar to what is described in linguistic studies. If more data are adopted and more dialectal features (initials and tones) are considered together, a good measurement of the

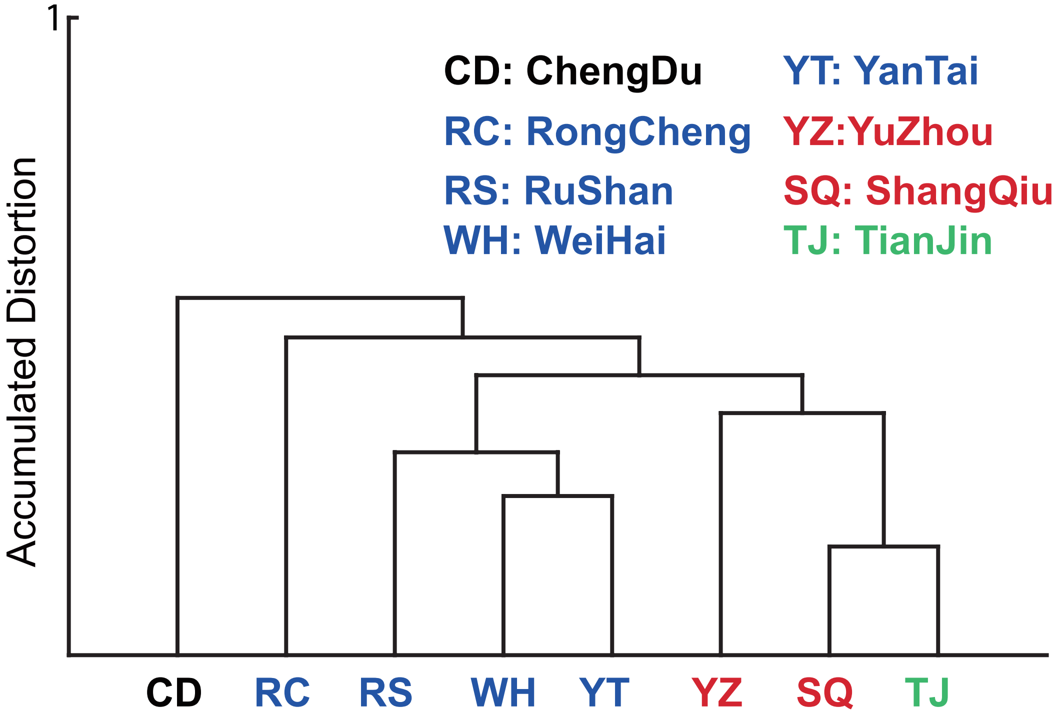


Figure 6.2: Sub-dialect-based speaker classification

acoustic distances among dialects can be obtained and it could be a good and objective proof of the study of linguists.

6.5 Conclusion

In this chapter, two classification experiments are carried out using sub-dialect data of Mandarin and the results show the structural method can extract the linguistic features from dialect pronunciations and classify speakers based on their dialects or sub-dialects. For the experiments, new data of 16 sub-dialect speakers of Mandarin are recorded in China. Then sub-dialect-based speaker classification experiment is carried out through building the sub-dialect pronunciation structures and the result shows that speakers are mainly classified based on their sub-dialects. After that, sub-sub-dialect based classification is carried out by calculating the distances between the sub-dialects among the hometowns and the result shows that the speakers can be well classified based on the dialects of their hometowns. In brief, all the experiments prove that our method can also work well in sub-dialect-based speaker classification.

Chapter 7

Verification of speaker-invariance using data of minimum speaker differences

7.1 Introduction

In last two chapters, dialect-based speaker classification experiment and sub-dialect-based speaker classification experiment are carried out and the results show that the structural pronunciation representation can extract the purely linguistic features and classify speakers based on their dialects or sub-dialects. However, about the result of sub-dialect-based speaker classification experiment, although the speakers from the same cities are all classified near to each other, speakers from BeiFang sub-dialect region are not classified together. After that, three possible reasons are given for it. Firstly, only the finals of the syllables are used in this experiment, but the initials are also considered as a very important factor in the linguistic classification of sub-dialects. Secondly, as the speakers are all graduate students and had been in BeiFang sub-dialect region for several years, their sub-dialect pronunciations may be already affected by other sub-dialects more or less. Thirdly, our classification experiment is carried out based on the acoustic features but some other features such as the articulatory, historical features of the dialects are also focused on in the traditional linguistic classification. Anyway, none of the above reasons can be proved, so we designed the following comparison experiments to prove that it is not the problem of our method.

In this chapter, two classification experiments are designed to prove that the structural method can classify speakers based on their dialects by extracting the speaker-invariant features. At the beginning, the data of 19 speakers from 10 different sub-dialect regions of 5 major dialect regions are recorded and dialect-based speaker classification experiment is carried out. Then a dialectologist mimicked the original data linguistically in her own voice and a new data set with minimum speaker differences is obtained. After that, a new classification experiment is carried out using the new mimicked data and the result is very similar to the result of above experiment. It means our structural method can classify speakers based on their results by extracting the speaker-invariant linguistic features.

7.2 Experiment design and data

7.2.1 Design of the experiments

In order to verify the speaker-invariance of our structural method, two comparison experiments are designed. First, using the dialect or sub-dialect data spoken by different speakers, a dialect-based speaker classification experiment can be carried out. Then if a new data set of multiple dialects with fixed speaker identity can be obtained, a similar classification experiment can be carried out and the result can be compared with the above one. If these two results are the same or very similar to each other, we can say that the speaker-invariance of our method is proved because the classification with fixed speaker identity will give us an ideal result of dialect classification.

7.2.2 Original dialect data

For this experiment, some new recordings were carried out at the NanKai University, China. 19 speakers joined our recordings. They belong to 10 different sub-dialect regions from 5 general dialect regions. Every speaker is given a speaker ID for the following experiments, and more information of them can be found in

Table 7.1: Detailed information of the dialect speakers

ID	Dialect	Sub-dialect	Hometown	Gender
M1	Min	QuanZhang	JiJang	F
M2	Min	QuanZhang	XiaMen	F
M3	Min	QuanZhang	QuanZhou	F
M4	Min	QuanZhang	XiaMen	M
Y1	Yue	GuangFu	FoShan	M
Y2	Yue	GuangFu	GuangZhou	F
Y3	Yue	GuangFu	FoShan	F
Y4	Yue	GuangFu	GuangZhou	F
H1	Hakka	NingLong	GanZhou	M
H2	Hakka	YuGui	XiuShui	M
H3	Hakka	TongGu	TongGu	F
H4	Hakka	TongGu	TongGu	F
X1	Xiang	LouShao	JiShou	F
X2	Xiang	ChangYi	Xiangtan	F
X3	Xiang	LouShao	ShaoYang	F
X4	Xiang	ChangYi	Xiangtan	F
G1	Gan	GuangChang	FuZhou	F
G2	Gan	LiYang	ShangGao	F
G3	Gan	GeYang	LePing	F

Table 7.1. The recording materials were a list of written characters in [59], which is used for checking the finals among different dialects by Chinese dialectologists. Some examples of these characters and the corresponding syllables and finals are listed in Table 6.2.

The recordings were carried out in quiet rooms with a supervisor. The recording equipments included a high quality microphone fixed on the table, a linear PCM recorder of Sony company. The data was recorded as monophony and the sample rate is 44 kHz. During the recording, every speaker was asked to read the selected characters in their native dialects three times. After the recording, the data was labeled phonetically and manually by students of linguistics. The final part of every syllable was modeled as a single Gaussian distribution under the acoustic conditions shown in Table 5.3 and the pronunciation structure for every speaker was built.

7.2.3 Linguistically mimicked data

If there is a Chinese dialectologist who can speak all these dialects, he/she can repeat the recorded utterances linguistically and which will gives us a new data set with a constant speaker identity. In fact, nobody can speak all the Chinese dialects. However, an experienced dialectologist can label the dialect data with IPA symbols and then read every transcript by looking at the symbols and listening to the original utterance at the same time. Then a new data set with constant speaker identity can be obtained. At last, a dialectologist from the Institute of Linguistics, Chinese Academy of Social Sciences finished this challenging work.

At the beginning, every syllable of the original data was transcribed to initial/final using IPA symbols. Then in a sound proof room, she read every syllable in her own voice according to the linguistic content after listening to the original utterance. After the recording, the new data set was checked at least twice by different linguists. By listening to the original utterance and the corresponding new one, the new mimicked utterance was ensured to be linguistically the same as the original one. Fig. 8.3 shows the spectrums of the original utterance of one syllable and the mimicked version. By this figure, we can find that the spectral features of the original data and the mimicked version are quite different.

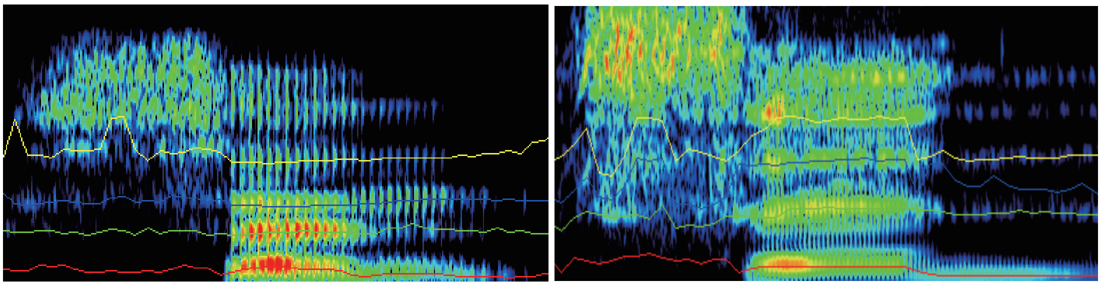


Figure 7.1: Spectrums of the original and mimicked data

At last, the new mimicked data was also labeled phonetically by students of linguistics. Then the final part of every syllable was modeled as a single Gaussian distribution under the acoustic conditions shown in Table 5.3. After that, for every speaker, the BDs between any pair of distributions were calculated and the dialect pronunciation structure was built.

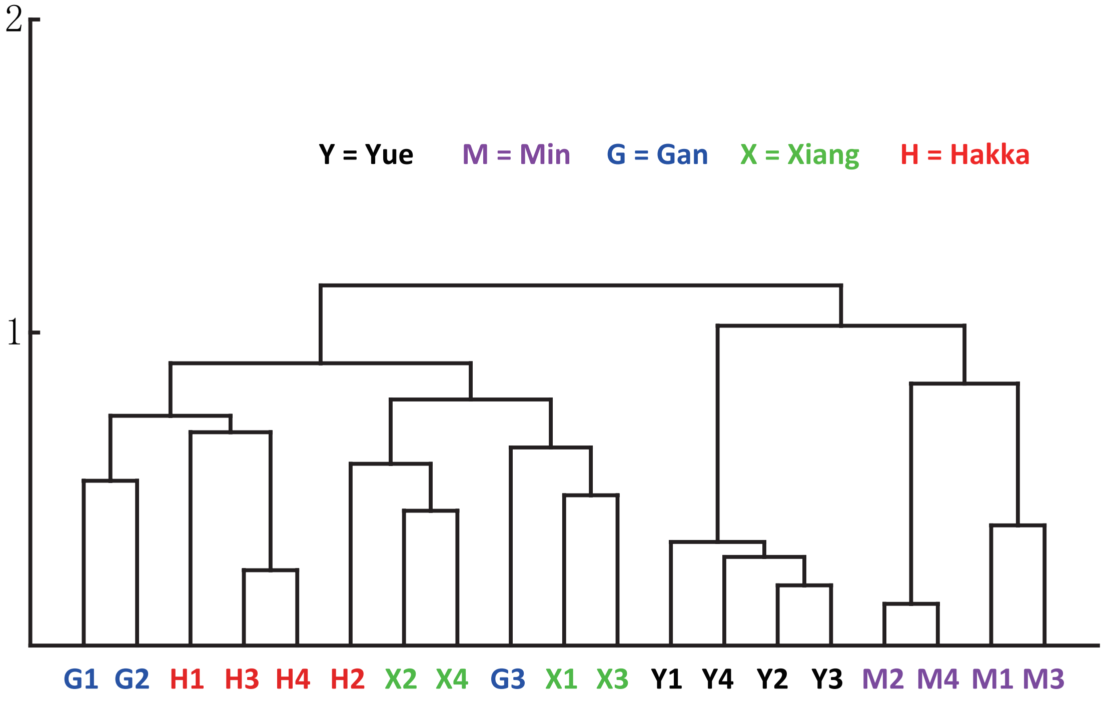


Figure 7.2: Result using the original dialect data

7.3 Comparison of the experimental results

After the pronunciation structures were built for the speakers, dialect-based speaker classification can be achieved by calculating the distances between their pronunciation structures. Using the original data spoken by different speakers, the distances between their dialects are obtained by calculating the distances between their structures and the classification result is shown in Fig. 7.2. Then using the mimicked data with constant speaker identity, similar experiment is carried out and the result is shown in Fig. 7.3. In both the figures, the structure of every speaker is represented by the speaker ID in Table 7.1 and different colors show different dialect regions.

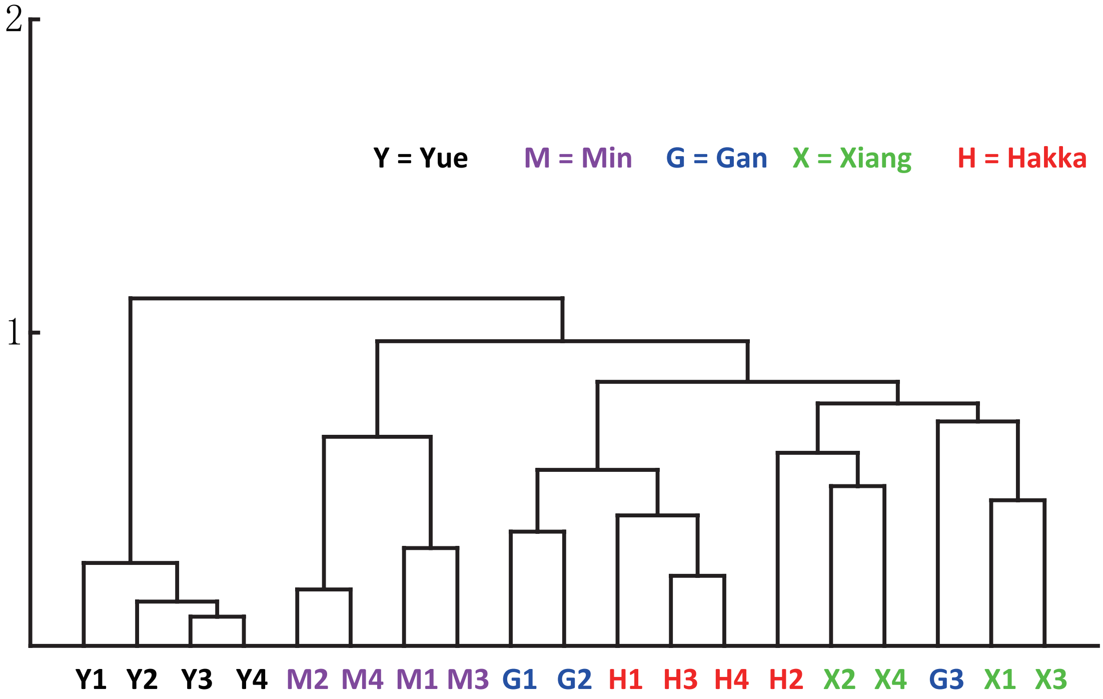


Figure 7.3: Result using the new mimicked data

In Fig. 7.2, we can focus on the speakers from Yue and Min dialect regions first, who are classified into a sub-tree on the right of this figure. Further, the speakers from Yue dialect and those from Min dialect are clustered to their sub-sub-trees. Meanwhile, about speakers from Hakka, Gan and Xiang, after checking the sub-dialect information of them in Table 7.1, we are able to determine the speakers from the same sub-dialect regions are all classified near to each other in the result. But looking at speakers G3 and H2, we have to admit that they are not completely clustered into different sub-sub-trees by their dialects.

In fact, the dialect regions of Hakka, Gan and Min are very near to each other geographically, genetically, phonologically. It is found that several sub-dialect regions of Hakka are located at the middle of Gan dialect region and the Xiang dialect regions are also very close to Gan dialect region geographically [28]. And about speaker G3 and H2, their data were checked by a dialectologist before she knew the result of this experiment. She found that the dialects of G3 and H2 were most different to other Gan speakers and Hakka speakers, respectively, and their three times' pronunciations of some characters were not very steady.

In Fig. 7.3, we can find that all speakers are classified into four large sub-trees: speakers from Yue and Min are classified into their individual sub-trees; speakers from Xiang are also classified into a sub-tree and speaker H2, G3 are also classified into this sub-tree as well; the left two Gan speakers and three Hakka speakers are clustered into a large sub-tree, which itself has two sub-sub-trees corresponding to Gan and Hakka separately.

By comparing the above two experimental results, it is found that they are very similar to each other. And only focusing on the Gan, Hakka and Xiang speakers, which is shown Fig. 7.4, we can find their positions are exactly the same. However, the positions of speakers Y2, Y3 and Y4 are somewhat different in the two results. In fact, this clustering tree is obtained using Ward's bottom-up method and the height means how different are the two groups. And in the results, it can found that both of the sub-trees of Yue speakers are very low and it means the their pronunciation are very similar to each other. Therefore, this difference between the two results is acceptable. So totally speaking, our method can extract the purely linguistic features by canceling the features of speaker differences.

7.4 Conclusion

In this chapter, two comparison experiments are carried out and the speaker-invariant feature of the structural method is proved. At the beginning, the dialect data of 19 speakers are recorded and dialect-based speaker classification experiment is carried out using the structural method. Then corresponding to the original data, a dialectologist mimicked them linguistically and a new data set with constant speaker identity are built. After that, a similar classification experiment is carried out using this new data data. By the results of the two

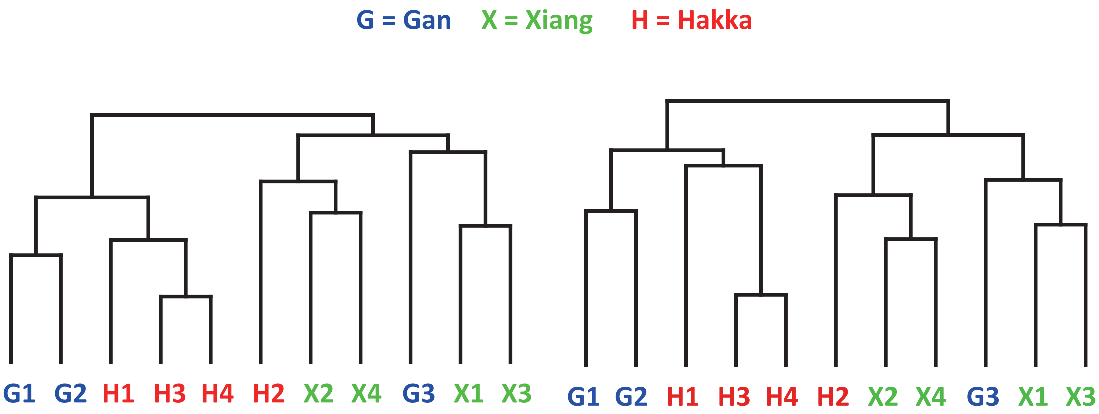


Figure 7.4: Comparison of the classification of Gan, Hakka and Xiang speakers

experiments, it is found that they are very similar to each other and the speaker-invariant feature of the structural method is proved.

Chapter 8

Comparison with spectral classification using data of maximum speaker differences

8.1 Introduction

In last chapter, the speaker-invariance of dialect-based speaker classification using structural method is verified by two comparison experiments. At the beginning, dialect-based speaker classification experiment is carried out using original dialect data spoken by 19 speakers. Then a new classification experiment is carried out using this new data set with minimum speaker differences. As these two results are very similar to each other, the speaker-invariance of our method is proved.

In this chapter, the structural method of dialect-based speaker classification is compared with the conventional method which is based on spectral comparison of speech events. At the beginning, corresponding to the original data and mimicked data in last chapter, data sets with maximum speaker differences are created using high-quality voice morphing techniques. Then using these data, classification experiments based on spectral comparisons are carried out. The results show that the classifications are affected greatly by the speaker features. After that, using these data, classification experiments are carried out using our structural method and the results show that these speakers are well classified by their dialects and it is not affected by the speaker differences. So our method is proved that unlike the conventional spectral comparison, it can extract the purely linguistic features and classify speakers based on their dialects.

8.2 Simulated data of tall and short speakers

8.2.1 Original and mimicked dialect data

For the following experiments, the original dialect data and the mimicked data introduced in last chapter are used here. The original dialect data includes the dialect data of 19 native dialect speakers who are from 10 different sub-dialect regions belonging to 5 major dialect regions (Min, Yue, Hakka, Xiang, Gan). The recording materials were more than 100 written characters in [59], which are used for checking the finals among different dialects by Chinese dialectologists. During the recording, every character was read three times. Then corresponding to the original data pronounced by different dialect speakers, every utterance was mimicked linguistically by one dialectologist in her own voice. So the mimicked data is a data set of multiply dialects but fixed speaker feature.

8.2.2 Data simulation using frequency warping

It is known that the vocal tract length of speaker is an important extra-linguistic feature and rotates a utterance trajectory in the cepstrum space [62]. Generally speaking, tall speaker always has long vocal tract, short speaker always has short vocal tract and the formants of utterances of speakers with long vocal tracts are lower than those of speakers with short vocal tracts. Using a frequency warping function, the utterances can be converted as if they are produced by the same speaker but with a much longer or shorter vocal tract. Frequency warping is characterized in the cepstral domain by multiplying c by matrix A (=$\{a_{ij}\}$).

$$\begin{aligned} a_{ij} &= \frac{1}{(j-1)!} \sum_{m=\max(0,j-i)}^{j} \binom{j}{m} \\ &\times \frac{(m+i-1)!}{(m+i-j)!} (-1)^{(m+i-j)} \alpha^{(2m+i-j)} \end{aligned} \tag{8.1}$$

where $|\alpha| \leq 1.0$, $m_0 = \max(0, j-i)$, and

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m). \end{cases}$$

When $\alpha < 0$, formants are modified to be lower and the vocal tract length longer. Otherwise, when $\alpha > 0$, formants are transformed to be higher and the vocal tract length shorter. And when the absolute value of α is bigger, the modification of the formants are bigger, which can be shown by Fig. 8.1.

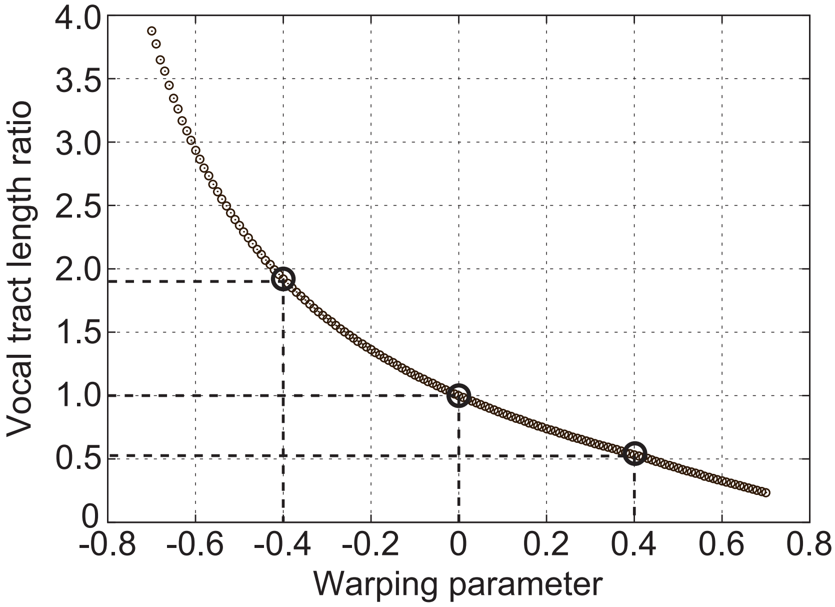


Figure 8.1: Relation between formants warping and α

The relation between the vocal tract length and the warping parameter α can be shown by Fig. 8.2. Here, considering the height of the world tallest adult and shortest adult, the original dialect and mimicked data were converted into a shorter version with $\alpha = 0.2$ and a taller version with $\alpha = -0.2$ using STRAIGHT [64]. The simulated data with $\alpha = 0.2$ sound like pronounced by a person with 1.3 times longer vocal tract, while the simulated data with $\alpha = -0.2$ sound like pronounced by a person with 0.7 times shorter vocal tract. Fig. 8.3 shows the spectrums of the same syllable produced by a Cantonese speaker and his two simulated versions. From left to right is the pseudo short speaker, original speaker and tall speaker.

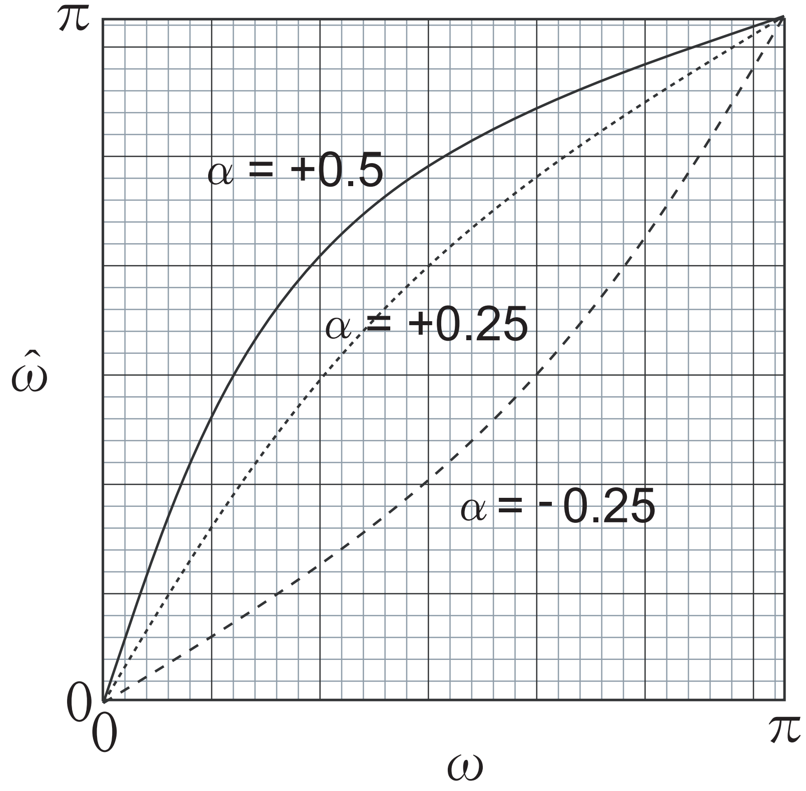


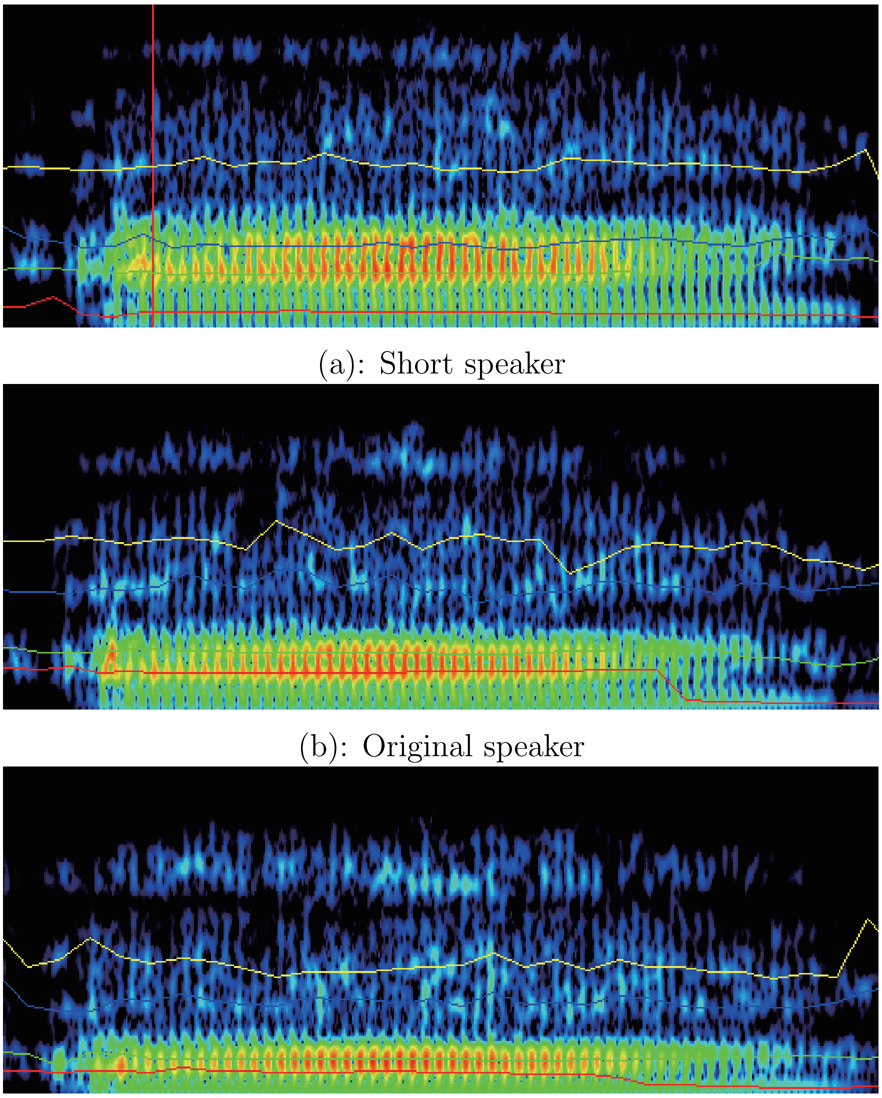
Figure 8.2: Relation between vocal tract length and α

8.3 Spectral classification using simulated data

In order to compare the dialect pronunciations of different speakers using our structural method, the distances between their pronunciation structures are calculated using the following formula:

$$D_1(S,T) = \sqrt{\frac{1}{M}\sum_{i<j}(S_{ij} - T_{ij})^2}. \qquad (8.2)$$

F_i^S is utterance i of speaker S and F_i^T is utterance i of speaker T. M means the number of the finals. In fact, in the conventional acoustic matching framework



(a): Short speaker

(b): Original speaker

(c): Tall speaker

Figure 8.3: Spectrums of short and tall speakers

such as DTW, for any pair of speech events, spectrums are directly compared between them. So if one want to calculate the distances between the dialects of two speakers based on the spectral comparison, the following formula can be used:

$$D_2(S,T) = \sqrt{\frac{1}{M}\sum_i BD(F_i^S, F_i^T)}. \tag{8.3}$$

Using the original dialect data and the simulated versions, a classification experiment is carried out by calculating the spectral distances between them using formula D_2. The result is shown by Fig. 8.4. The speaker IDs and the colors are the same meanings as they are in Table 7.1, while the ID with top bar means the simulated taller speaker and ID with under bar means the simulated shorter speaker. In this figure, speakers are classified into three big sub-trees according to their heights. It is found that in each sub-tree, the classification is affected by the speaker features and speakers are not classified by their dialects at all.

Then using the new mimicked data and the simulated versions, a classification experiment like above one is carried out and the result is shown by Fig. 8.5. The speaker IDs and the colors are the same meanings as they are in Fig. 8.4. Then by this figure, it is found that the speakers are also classified into three big sub-trees according to their heights. But in each sub-tree, the speakers from the same dialect regions are mainly classified near to each other, not like they are in Fig. 8.4. It is because that this result is carried out using the mimicked data and the speaker differences are already removed manually.

8.4 Structural classification using simulated data

Using the original dialect data and the simulated taller and shorter versions, dialect pronunciation structures are built and speakers are classified based on the distances between them. The result is shown by Fig. 8.6. After that, using the mimicked data and the simulated taller and shorter version, a similar experiment is carried out and the result is shown in Fig. 8.7. In both the results, the speaker IDs and the colors are the same meanings as they are in Table 7.1, while the ID with top bar means the simulated taller speakers and one with under bar means the simulated shorter speakers.

In Fig. 8.6, we can see that all the speakers are classified by their dialects and the result is not affected by the heights of the speakers. Every speaker and his/her simulated taller and shorter versions are all classified near to each other, because the dialectal features of their utterances are the same. In this result, all the speakers are classified into four sub-trees. Yue and Min speakers are classified into their individual sub-trees. Speakers from Xiang are classified into a sub-tree and speakers H2, G3 are also classified into this sub-tree as well. The left two Gan speakers, G1 and G2, and three Hakka speakers, H1, H3 and H4, are clustered into a sub-tree and they have two sub-sub-trees corresponding to Gan and Hakka separately. If we just focus on the dialects of the speakers, we can find this classification result is exactly the same as the result in Fig. 7.2 which is obtained only using the original dialect data.

Fig. 8.7 is obtained using the mimicked data and the simulated versions. The mimicked data has the constant speaker identity. The speaker features of the simulated versions are quite different to each other. In Fig. 8.7, it is found again that all the speakers are classified by their dialects and the simulated speakers are all classified near to the corresponding original ones separately: Yue and Min speakers are classified into a sub-tree; Xiang speakers are classified into a sub-tree together with G3 and H3; The left Gan and Hakka speakers are classified into one sub-tree.

The above two results using experimental data with quite different speaker features show that we can find the structural method still work very well. Speakers are classified by their dialects and the results are not affected by speaker features at all. By comparing these results with the results obtained in last section, it is further proved that unlike the classification using conventional spectral comparison, our structural method can extract the purely speaker-invariant dialect features from speech.

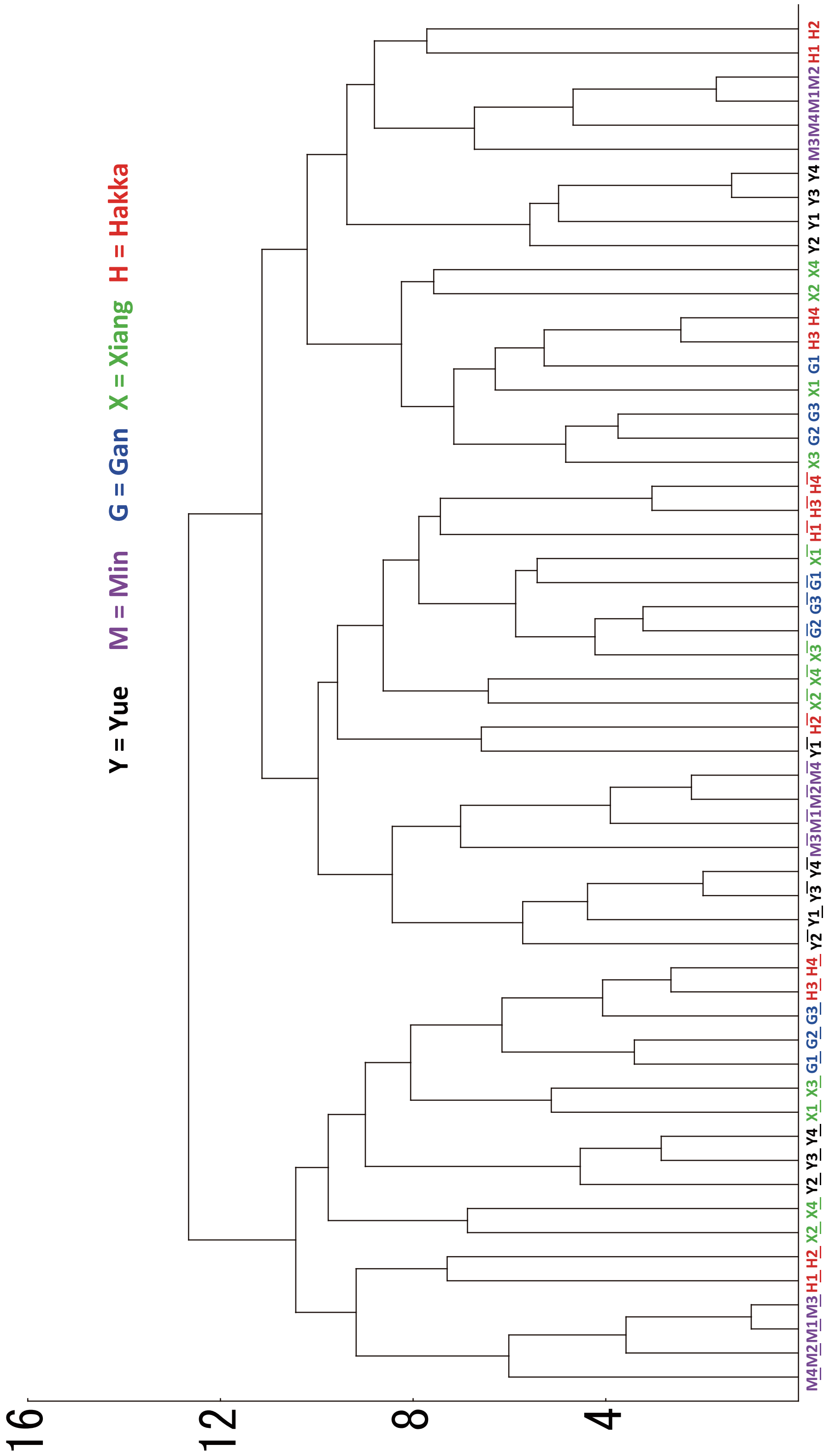


Figure 8.4: Speaker classification based on spectral comparison using the simulated data of the original data

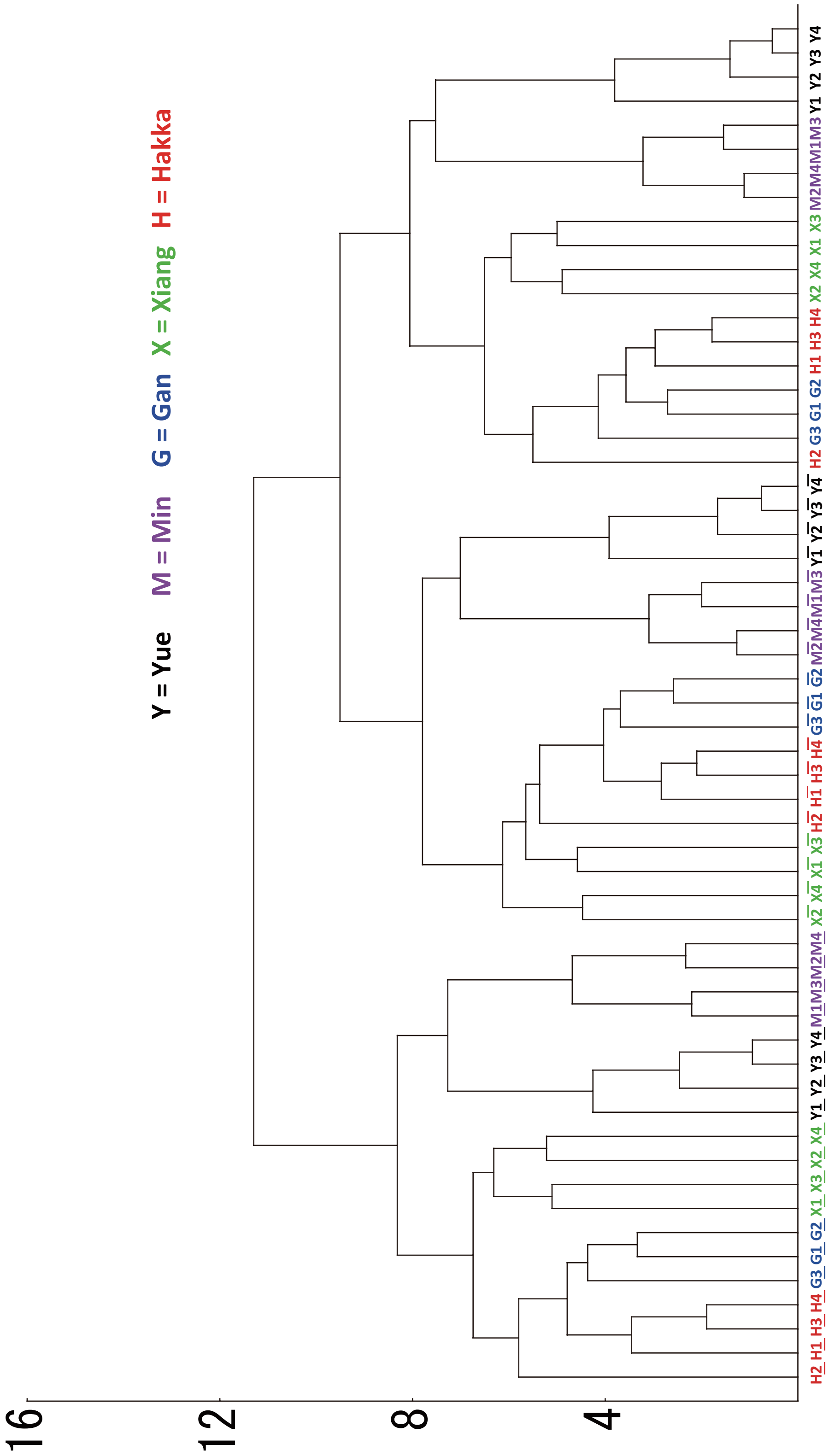


Figure 8.5: Speaker classification based on spectral comparison using the simulated data of the mimicked data

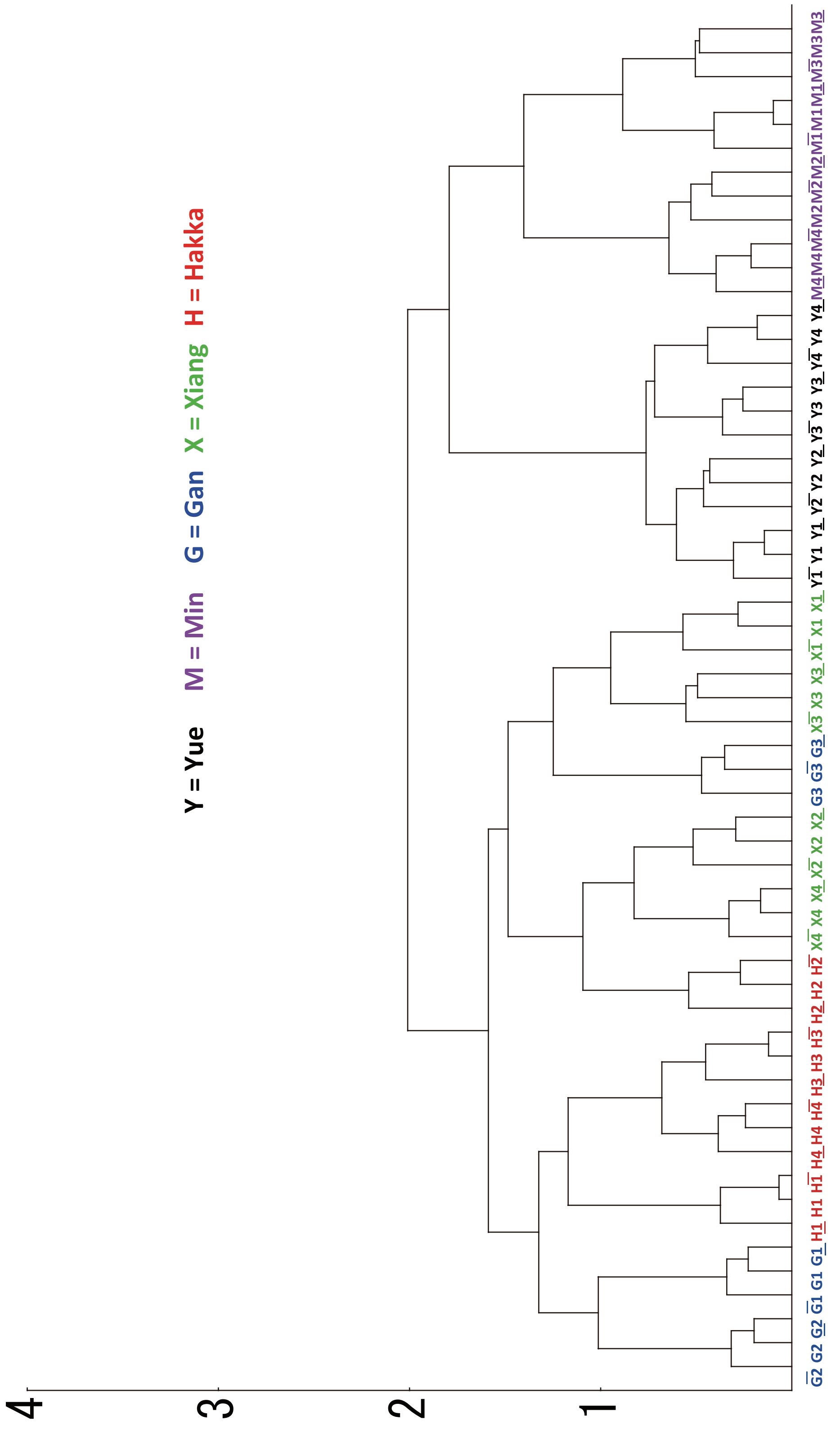


Figure 8.6: Speaker classification based on structural comparison using the simulated data of the original data

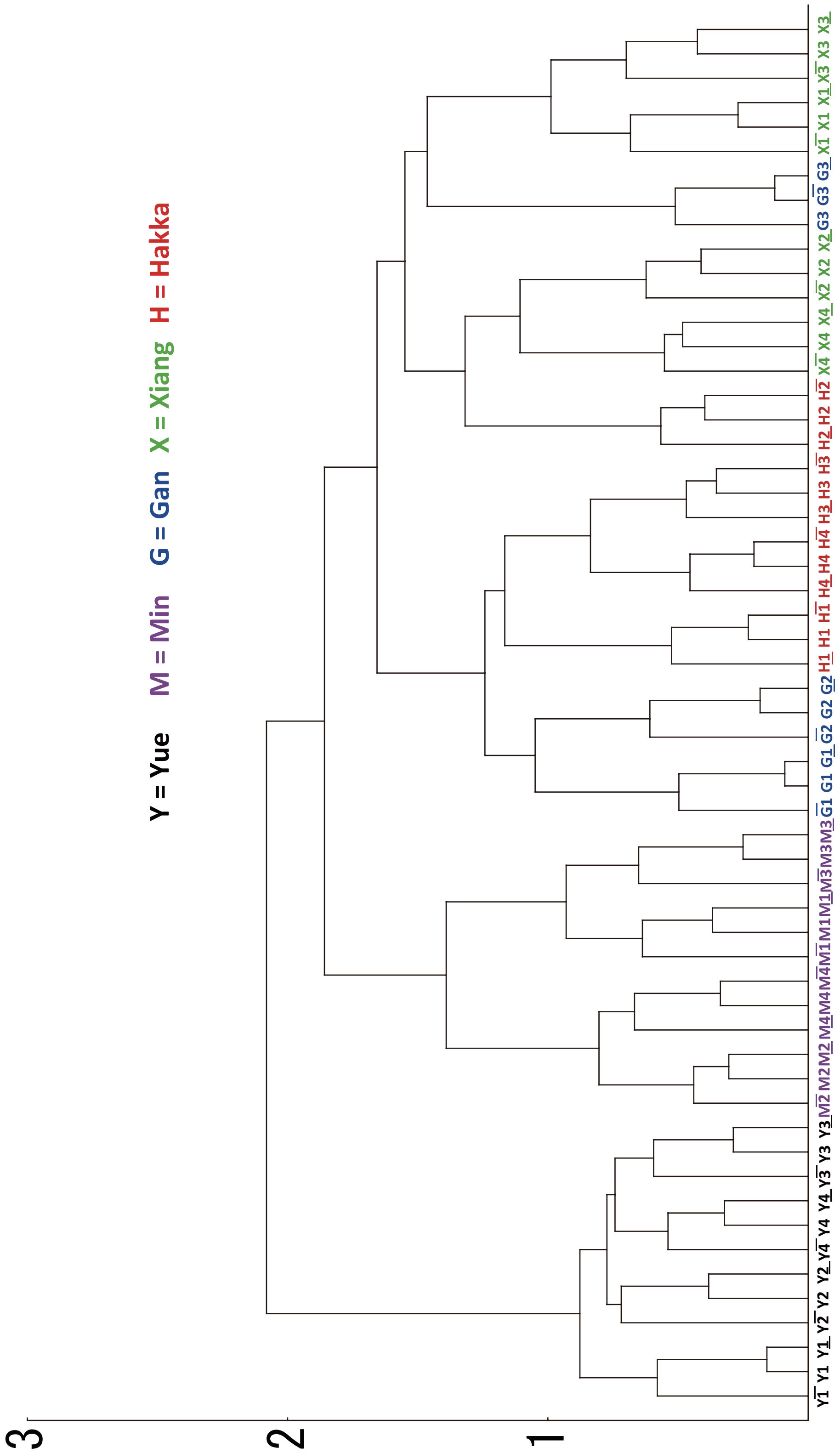


Figure 8.7: Speaker classification based on structural comparison using the simulated data of the mimicked data

8.5 Conclusion

In this chapter, the structural method of dialect-based speaker classification is compared with the conventional classification method based on spectral comparison using dialect data with maximum speaker differences. At the beginning, corresponding to the original dialect data and mimicked data, new data sets with maximum speaker differences are created using high-quality voice morphing techniques. Then using these data, classification experiments based on spectral comparisons and structural comparisons are carried out and the results show that unlike the method of spectral comparison, the structural method can classify these speakers based on their dialects by extracting the speaker-invariant purely linguistic features.

Chapter 9

More applications of the structural method

9.1 Introduction

Through the above chapters, it is proved by several different experiments that the structural representation of Chinese dialects can classify speakers based on their dialects or sub-dialects by extracting the purely linguistic features from speech after removing the extra-linguistic features like speaker features. At the beginning, using some dialect data and sub-dialect data, two experiments, dialect-based speaker classification experiment and sub-dialect-based speaker classification experiment, are carried out using our structural method separately and the results show that these speakers can be well classified by their dialects or sub-dialects. After that, comparison classification experiments are carried out using the original dialect data spoken by different speakers and the corresponding linguistically mimicked data with constant speaker identity separately. As the final results are almost the same to each other, our method is proved that it is invariant to speaker features. At last, our structural method is compared with the conventional method of spectral comparison using dialect data with maximum speaker differences. Using a frequency warping function, data set with maximum speaker differences are converted as if they are produced by the same speaker but with a much longer or shorter vocal tract. The result of the spectral comparison shows that the classification is affected by speaker features greatly and the speakers are mainly classified by their heights, while our structural method can classify speakers based on their dialects and it is independent to the speakers features.

In this chapter, the structural method is further applied to several different applications and several preliminary experiments are carried out. At the beginning, for every speaker, his/her dialect utterances are compared with the corresponding standard Mandarin utterances and a similarity order between these utterances can be obtained. By the results, it is found that the similarity orders of the speakers from the same regions are very similar. Further, we try to apply this method method to pronunciation assessment of accented Mandarin. At the beginning, every accented Mandarin is given a structural score using the structural method. And these utterances are evaluated manually by trained Mandarin speakers and a manual evaluated score is given. Also, these utterances are evaluated by a Mandarin recognizer and another score is given. At last, the results of these three evaluation methods are compared and the results are discussed.

9.2 Estimation of utterances similarity orders

9.2.1 Comparison of individual utterances

About the dialect-based speaker classification experiments, the dialects of different speakers are compared by calculating the distances between their whole dialect pronunciation structures. In fact, this procedure can be decomposed into the comparison of individual utterances using the following formula:

$$d(A, B, v) = \sum_{v} |A_{vi} - B_{vi}|, \tag{9.1}$$

where A and B mean the matrices of two compared pronunciation structures, v means the utterance to be compared. If we take A as the pronunciation structure of a dialect speaker and B as the pronunciation structure of a standard Mandarin speaker, the utterance of the largest d means this utterance is the most different one comparing all the utterances of this dialect speaker with the standard Mandarin speaker. The utterance of the smallest d means this utterances is the most similar one comparing the utterances of these two speakers. Then using this method, the utterances of different dialect speakers can be compared with standard Mandarin speaker and the similarity orders can be obtained.

9.2.2 Similarity estimating experiments

For this experiment of estimating the similarity orders of utterances, the data of four speakers are adopted. One is a male Min dialect speaker, one is a female Min dialect speaker, one is a male standard Mandarin speaker and the last one is a female standard Mandarin speaker. The recording materials are the characters in Table. 9.1. During the recording, every character was read three times. After that, the recording data was labeled manually and the syllable part was cut. Then the distribution is calculated for every syllable and the pronunciation structure is built for every speaker using the BDs. After that, the pronunciation structures of the Min dialect speakers are compared with the standard Mandarin speakers and the utterance similarity orders between them can be calculated using $d(A, B, v)$.

Table 9.1: Selected characters for dialect pronunciation assessment

Characters	辣、架、蛇、落、野、 月、資、知、耳、第、虚、 木、北、桂、桃、薬、流、 三、根、温、林、減、園、 隣、党、講、東、瓮、瓊
Syllables	/la/,/jia/,/she/,/luo/,/ye/, /yue/,/zi/,/zhi/,/er/,/di/,/xu/, /mu/,/bei/,/gui/,/tao/,/yao/,/liu/, /san/,/gen/,/wen/,/lin/,/jian/,/yuan/, /lin/,/dang/,/jiang/,/dong/,/weng/,/qiong/

After the utterance similarity orders between the Min dialect speakers and the standard Mandarin speakers are calculated, they can be shown by the following figures: Fig. 9.1 shows the utterance similarity order between the female Min dialect speaker and the female Mandarin speaker. Fig. 9.2 shows the similarity order of the utterances between the male Min dialect speaker and the female

Mandarin speaker. Fig. 9.3 shows the similarity order of the utterances between the female Min dialect speaker and the male Mandarin speaker. Fig. 9.4 shows the similarity order of the utterances between the male Min dialect speaker and the male Mandarin speaker. In all these figures, the X-axis means the utterances and the Y-axis represents d of the utterances. From left to right on the X-axis, d is reducing. It means the dissimilarity between the dialect utterances and Mandarin utterances are reducing.

By comparing these figures, it is found the utterance similarity orders between Min dialect speakers and Mandarin speakers are very similar to each other. For example, by the results shown in Fig. 9.1 and Fig. 9.2, which are obtained by comparing the utterances of two Min dialect speakers of different genders with one female Mandarin speaker, it is found the first several characters and the last several characters are the same. So it means the dialect utterances of these two Min speakers are very similar to each other. In fact, these two Min speakers are born and brought up at the same city and their utterances of these characters are supposed to be very similar to each other. Then by comparing Fig. 9.3 with Fig. 9.4, which are obtained by comparing the utterances of the two Min dialect speakers with the male Mandarin speaker, very similar conclusion can be obtained. And by comparing these results together, some common features can be found. For example, the utterances of the six red characters /辣, 林, 耳, 桂, 資, 知/ are all located at the left edge in all the results. It means the Min pronunciation of these characters are very different to their Mandarin pronunciations. Further, this result also show high independent to extra-linguistic features, such as the gender of the speakers. So, this method can be applied to comparing the dialect utterances of any speaker with the standard Mandarin utterances and the result is not affected by extra-linguistic features.

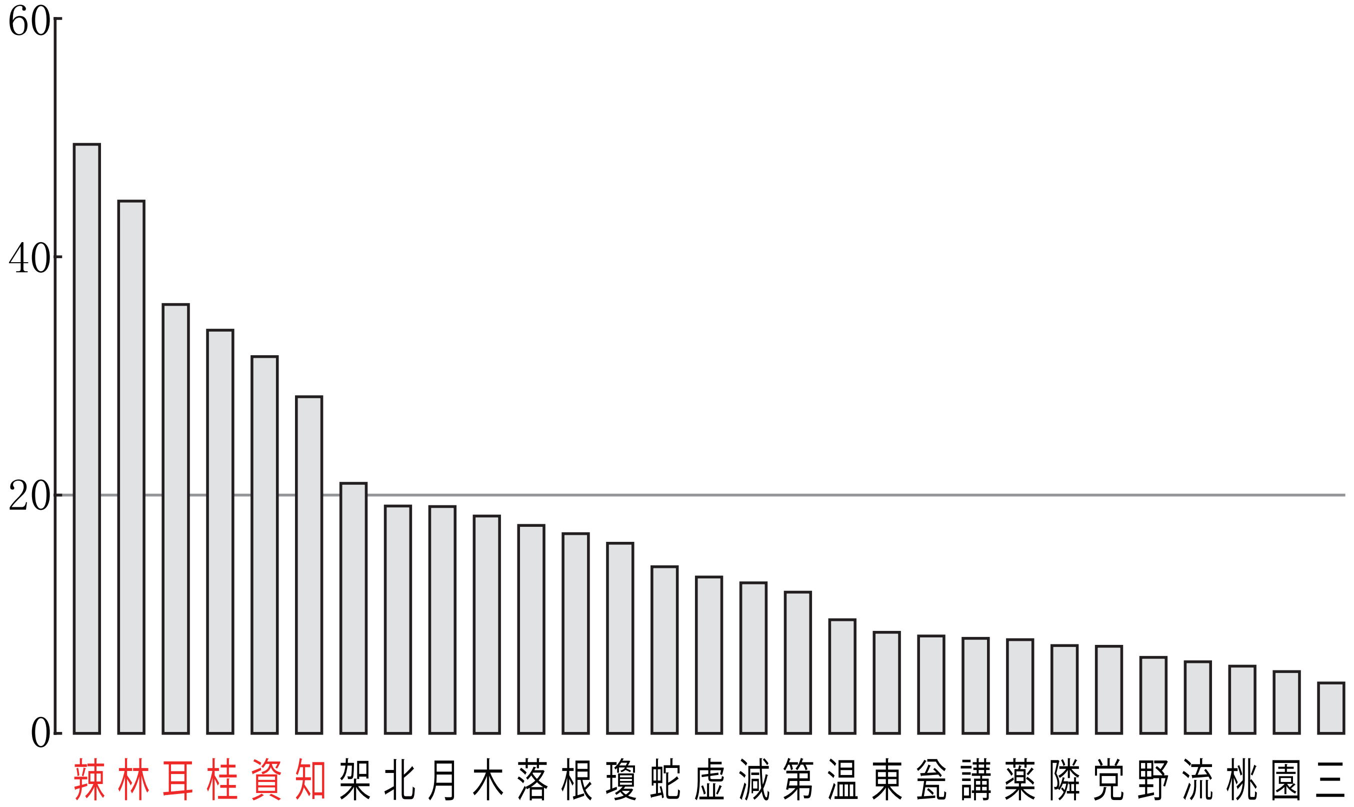


Figure 9.1: Utterance similarity order between a female Min speaker and a female Mandarin speaker

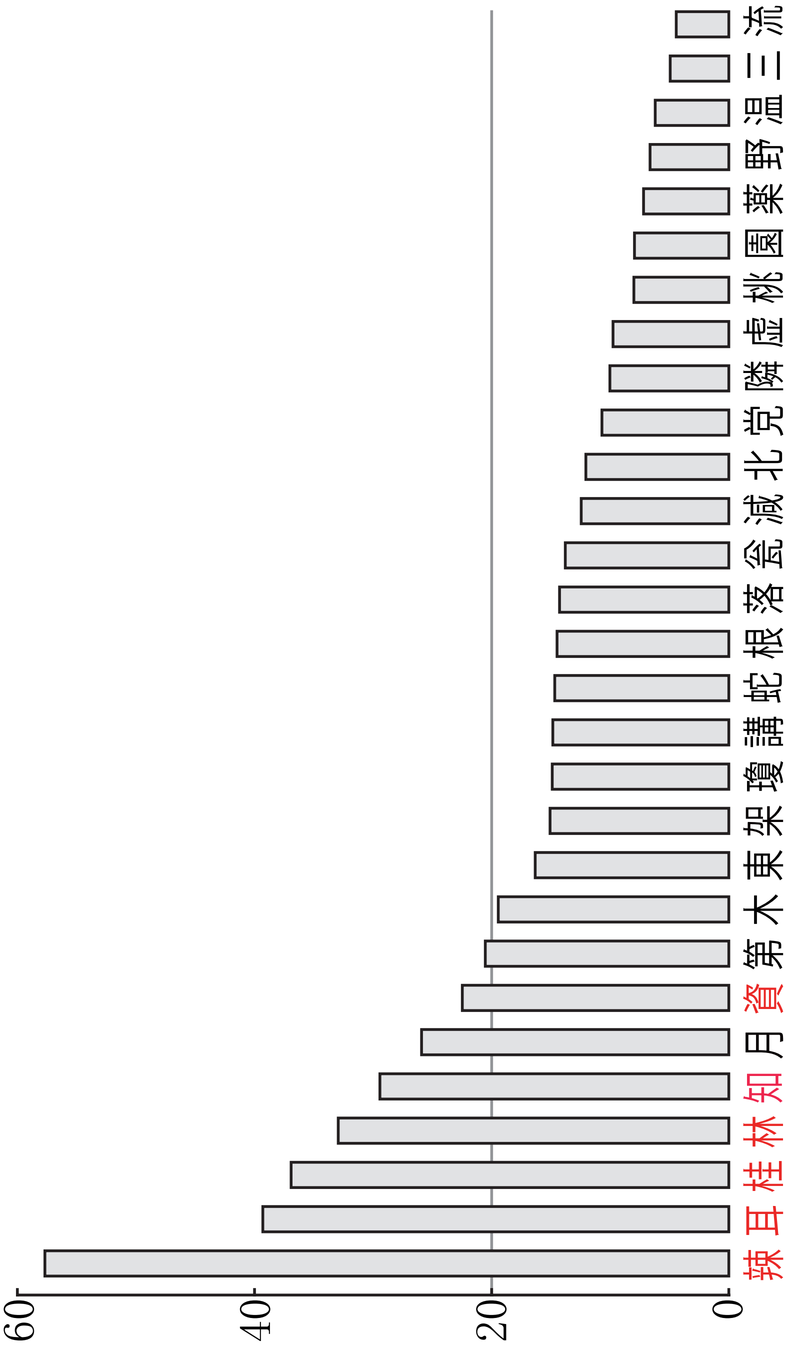


Figure 9.2: Utterance similarity order between a male Min speaker and the female Mandarin speaker

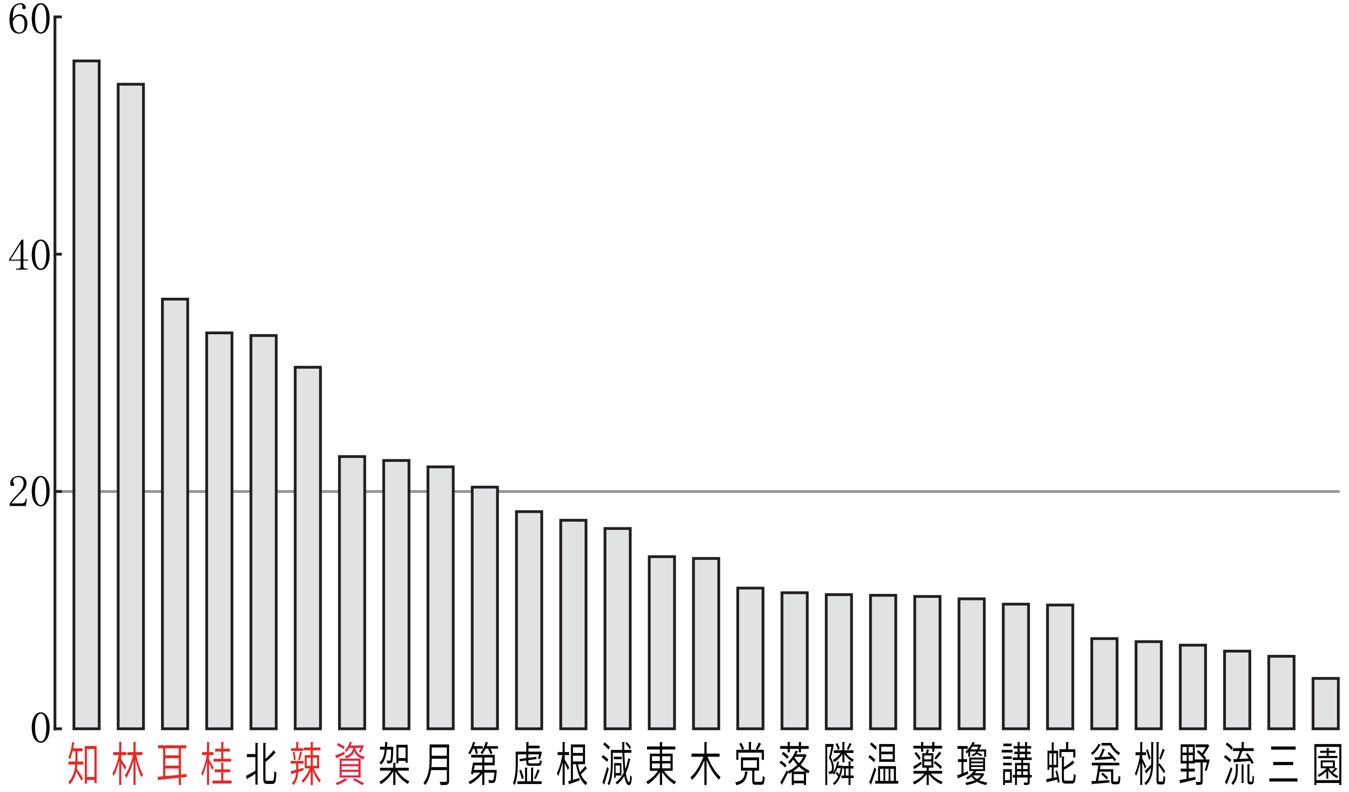


Figure 9.3: Utterance similarity order between a female Min speaker and a male Mandarin speaker

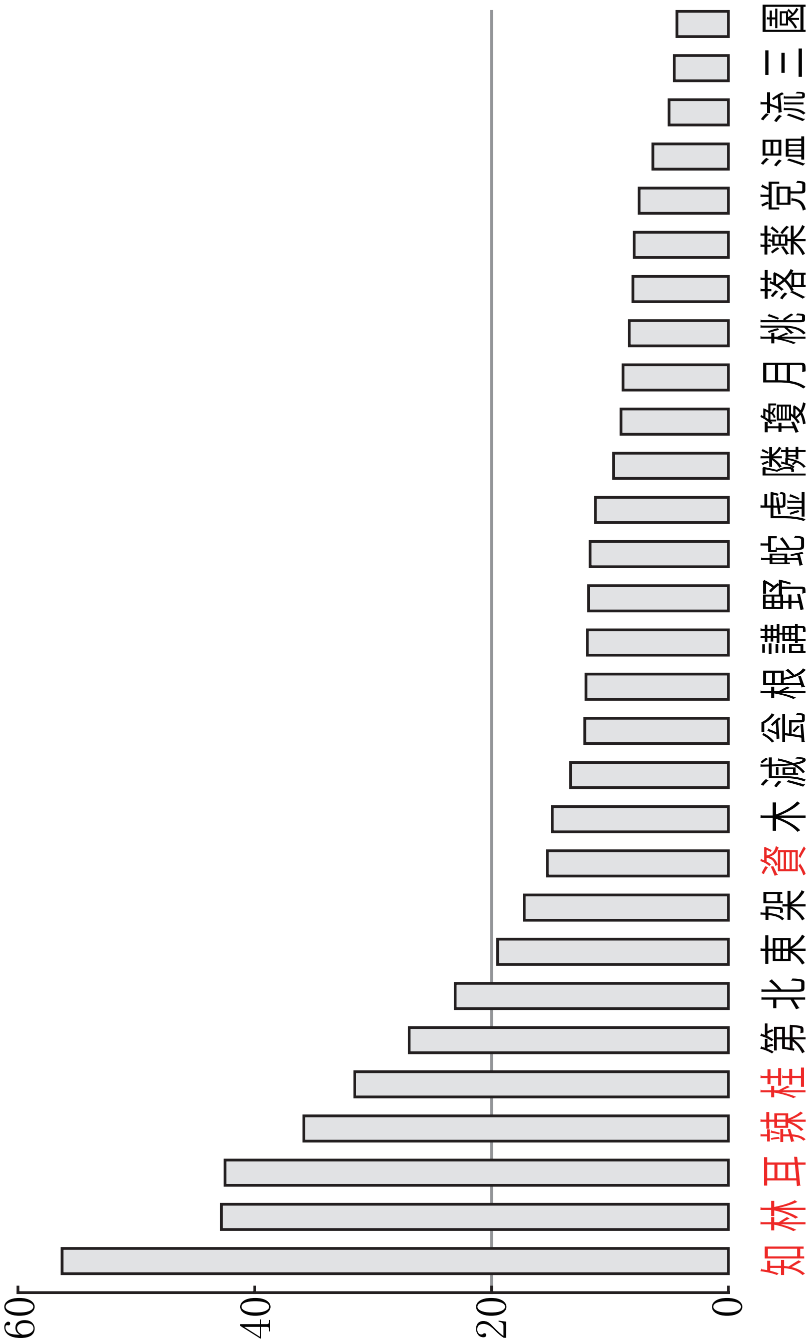


Figure 9.4: Utterance similarity order between a male Min speaker and a male Mandarin speaker

9.3 Accented Mandarin assessment

9.3.1 Structural evaluation score

In last section, the structural method is applied to calculating the utterance similarity between dialect speaker and standard Mandarin speaker. In this section, this structural method is further applied to pronunciation assessment of accented Mandarin. Using the accented Mandarin pronunciations of dialect speakers, new pronunciation structures can be built. Then these pronunciation structures can be compared with a standard Mandarin structure and a structural evaluation score can be calculated for individual utterances using $d(A, B, v)$.

In order to verify this proposal by experiment, new accented Mandarin data were recorded. For the speakers in Table 7.1, their accented Mandarin pronunciations were recorded together with the data of four standard Mandarin speakers. They were asked to read the characters used in Section 7.2.2 and some examples of them are shown by Table 6.2. The recordings were carried out in a quiet room in China. After that, the data was labeled manually and the final parts were analyzed under the acoustic conditions in Table 5.3. Each final was modeled as a diagonal Gaussian distribution and the parameter estimation was done for Gaussian modeling using MAP (Maximum A Posteriori) criterion. Then for the accented Mandarin and standard Mandarin speakers, their pronunciation structures are built. After that, the utterance of every accented Mandarin speaker is compared to standard Mandarin and the related distance is calculated as the structural score using $d(A, B, v)$.

9.3.2 Manual evaluation score

Meanwhile, all the accented Mandarin utterance was evaluated manually by the staffs in iFlytek[1] and their daily job is Mandarin pronunciation assessment. After the utterances are listened by the staffs, the initial, final and tone of every syllable utterances is rated by 0,1 and 2. 0 means the pronunciation is correct, 2 means the pronunciation is completely wrong and 1 means the pronunciation has a flaw. And for every speaker, the numbers of the utterances rated by score 1 and 2 are

[1] http://www.iflytek.com

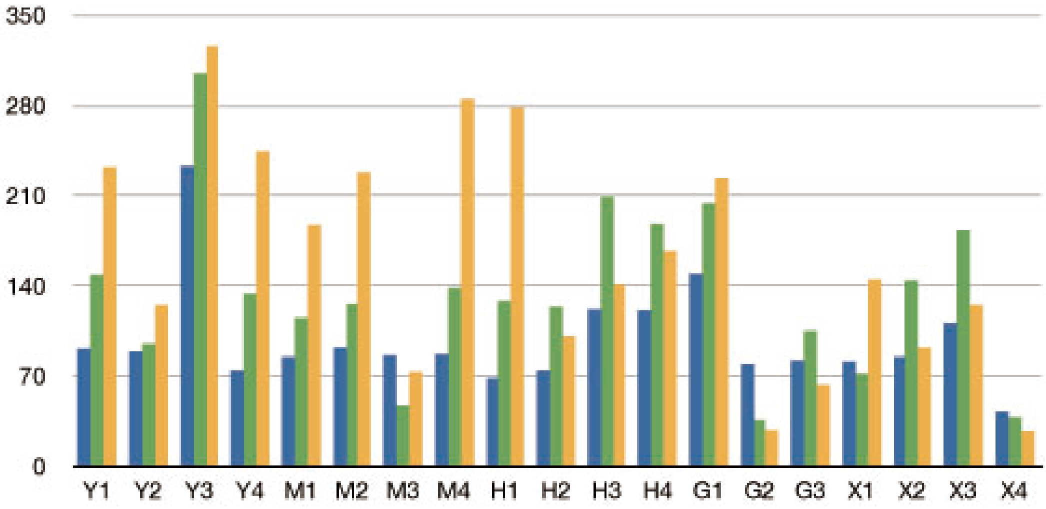


Figure 9.5: Numbers of utterances evaluated with score 1 and 2

shown by Fig. 9.5, the numbers of the utterances rated by score 1 are shown by Fig. 9.6 and the utterance numbers of score 2 are shown by Fig. 9.7. In these figures, the blue bars mean the numbers of initials, green bars mean the numbers of finals and orange bars mean the numbers of tones. So by comparing these figures, it is found that accented Mandarin pronunciation of these speakers are quite different. The Mandarin pronunciation of speaker Y2, G2, G3, X4 are better that other speakers, totally speaking. Then focus on Fig. 9.6, the numbers of utterances having a flaw, it is found many speakers made some mistakes about the tones, less mistake are made about the initials, the least is the finals. About Fig. 9.7, it is found that the numbers are mainly less than the pronunciations with flaw in Fig. 9.6. Some pronunciations of several speakers are completely wrong, such as speaker Y3 , H3, H4, G1.

9.3.3 Compare with manual evaluation scores

Here, for every utterance, the manual evaluation score is compared with the structural evaluation score. After the accented Mandarin pronunciation structure is built for every speaker, his/her utterances are compared with the correspond-

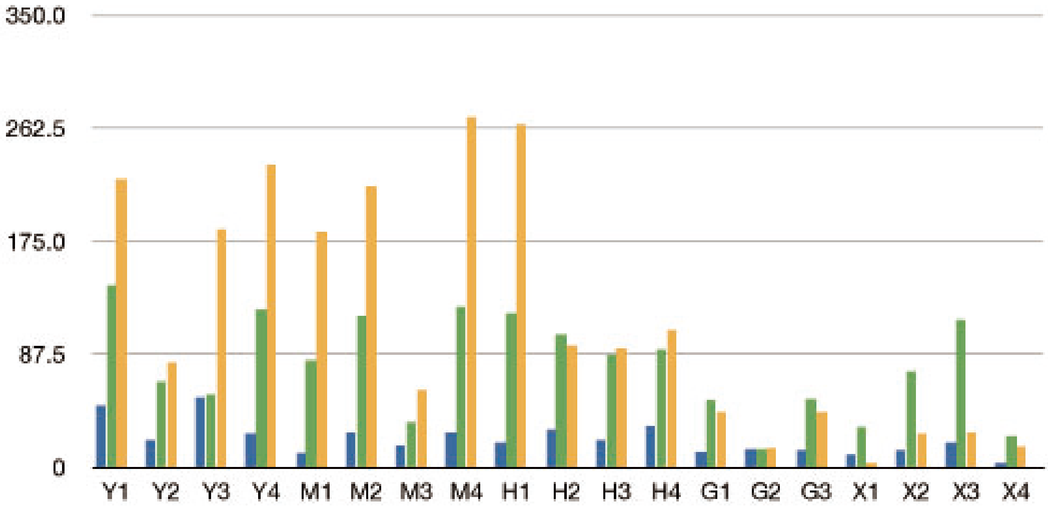


Figure 9.6: Numbers of utterances evaluated with score 1

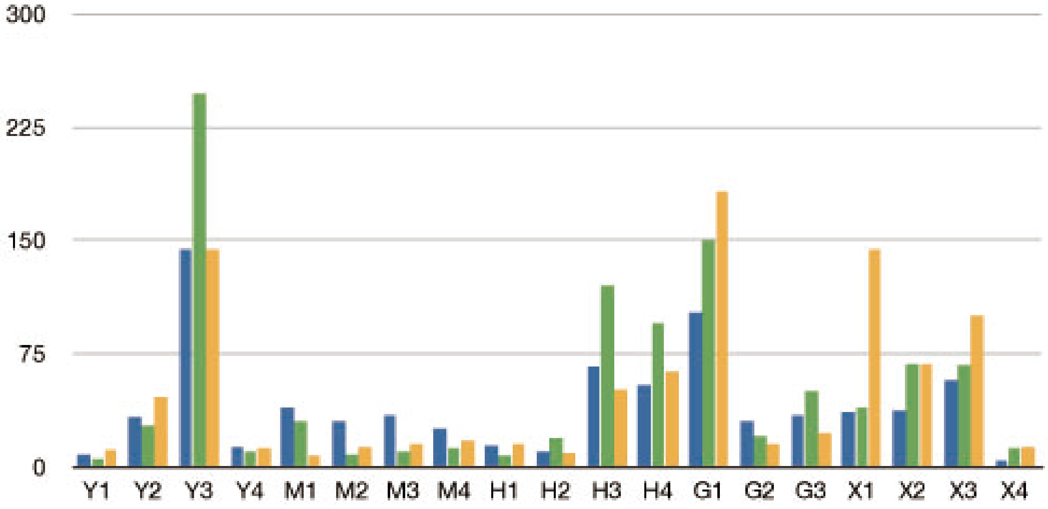


Figure 9.7: Numbers of utterances evaluated with score 2

ing utterances of one standard Mandarin speaker using the $d(A, B, v)$. Then a

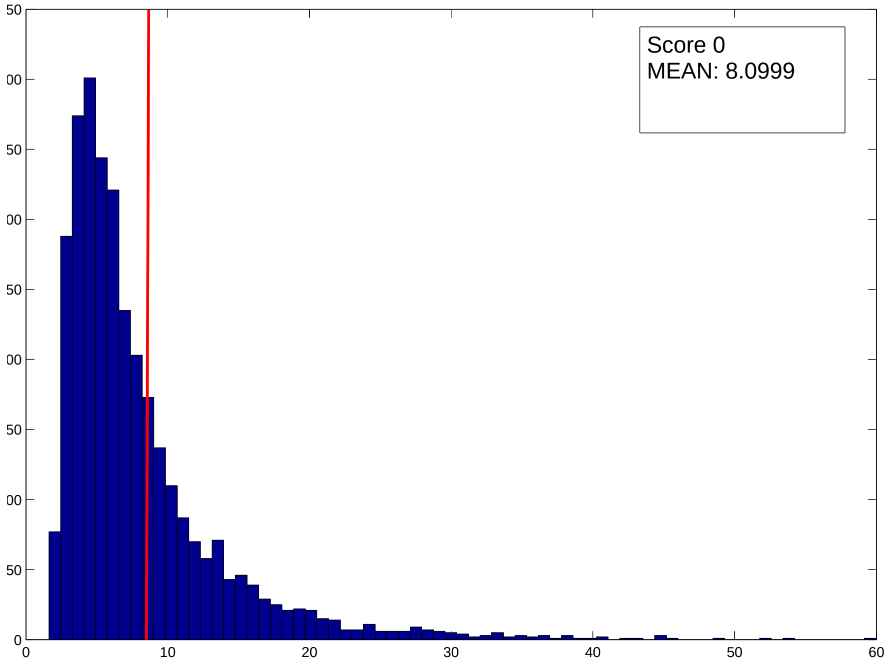


Figure 9.8: Distribution of structure distances (Score 0)

structural score vector can be obtained by the following formula

$$V(A, B) = (d(A, B, v_1), d(A, B, v_2), ..., d(A, B, v_n)). \tag{9.2}$$

A means the structure of accented Mandarin speaker, B means the structure of standard Mandarin speaker, v means the utterance and n means the total number of the utterance events. For every accented Mandarin speaker, his/her utterance structure is compared with four standard Mandarin speakers and four structural score vectors are obtained. After that, these structural score vectors are connected and a large vector V_s is obtained. Then, using the corresponding manual evaluated scores (0, 1, 2), a large manual evaluated score vector V_m is obtained. The correlation coefficient between these two scores are calculated as 0.17. After that, several normalization techniques are used to calculate the correlation coefficient between these two scores. For example, an average structural score vector

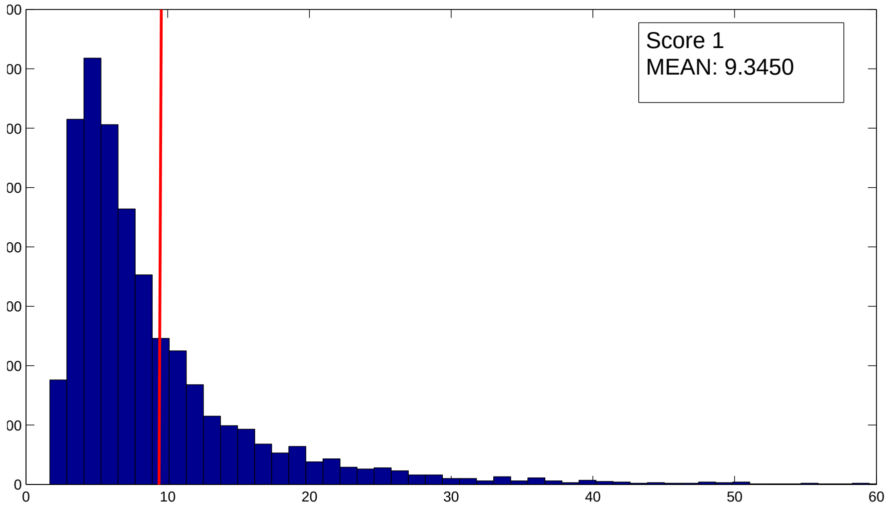


Figure 9.9: Distribution of structure distances (Score 1)

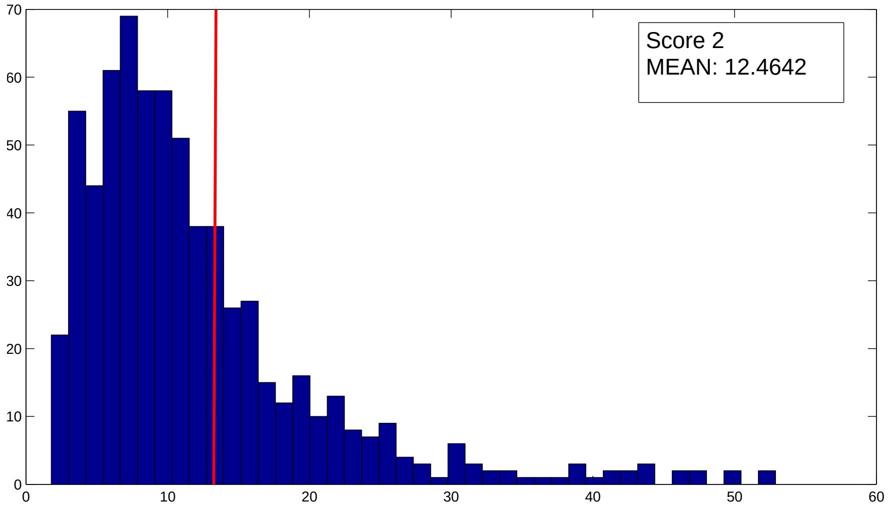


Figure 9.10: Distribution of structure distances (Score 2)

is calculated for every accented Mandarin speaker using the structural vectors obtained by the comparison with four standard Mandarin speakers. The new correlation coefficients between these two averaged evaluation scores are about 0.20 and it is not so satisfactory. After that, a different method is adopted to show the relationship of these two scores. The finals are selected separately by the manual evaluated score 0,1,2. Then the corresponding structural scores for these finals are shown by Fig. 9.8, Fig. 9.9 and Fig. 9.10 together with their mean. The X-axis means structural score and the Y-axis means the utterance numbers. By comparing these three figures, the mean structural scores of Fig. 9.10 is higher that Fig. 9.9 and which is higher than Fig. 9.8. So it means that wrong utterances get higher structural evaluation scores, just as we supposed.

9.3.4 Compare with results of speech recognition

In fact, about the above manual evaluation of the accented Mandarin utterances, the scores may be different to different listeners. So another method is designed to verify our proposal of structural evaluation. At the beginning, a phoneme based HMM is trained and a standard Mandarin recognizer is built. Then the accented Mandarin utterances are recognized. The recognition rate for every speaker is shown by Fig. 9.11. In this figure, the blue bars mean the recognition rates of the accented Mandarin. Then by comparing this figure with Fig. 9.5, we can find that the speakers get low scores in Fig. 9.5 mainly get high recognition rate in Fig. 9.11, such as Y2, M3, G2 and so on.

9.3.5 Pronunciation assessment using sub-structure

In fact, the pronunciation structures of accented Mandarin used above are built with pronunciations of more than 100 characters. Because these utterances are selected trying to cover the phonological differences among all the dialects, many of their accented Mandarin utterances are very similar to each other. So the BDs between these utterances are very short and many edges of the built structures are very short. Then when this structure is compared with the structures of standard Mandarin, the calculated related distances for some utterances are not very accurate. In order to avoid this problem, sub-structures are built by reducing the size of these structures using utterance selections.

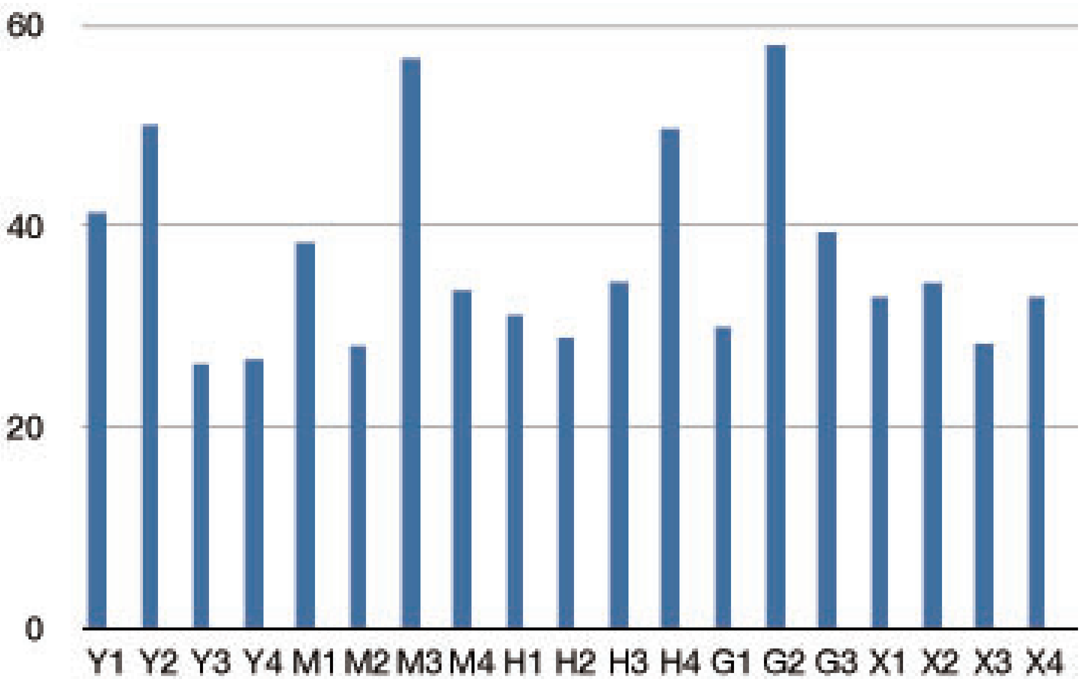


Figure 9.11: Recognition rate of accented Mandarin and Mandarin

Fig.9.12 shows how to build sub-structures to represent the accented Mandarin pronunciation of speakers and the procedures is the same as [65]. After the pronunciation structure is built for accented Mandarin speaker and standard Mandarin speaker using BDs among their utterances, sub-structure can be extracted by selecting a sub-set of their utterances. Then the accented Mandarin pronunciation can be assessed by comparing its sub-structures with the sub-structure of standard Mandarin speaker.

Using the utterance of 19 accented Mandarin speakers and standard Mandarin speaker, the following experiments are carried out. At the beginning, sub-structure is built by adding utterances and the adding sequence is fixed using cross validation with 18 speakers for training and one left speaker for testing. Then a new structural evaluation score is given for every utterance and a score vector is obtained for all the utterances of this sub-structure. After that, the correlation coefficient between this structural score vector and the corresponding

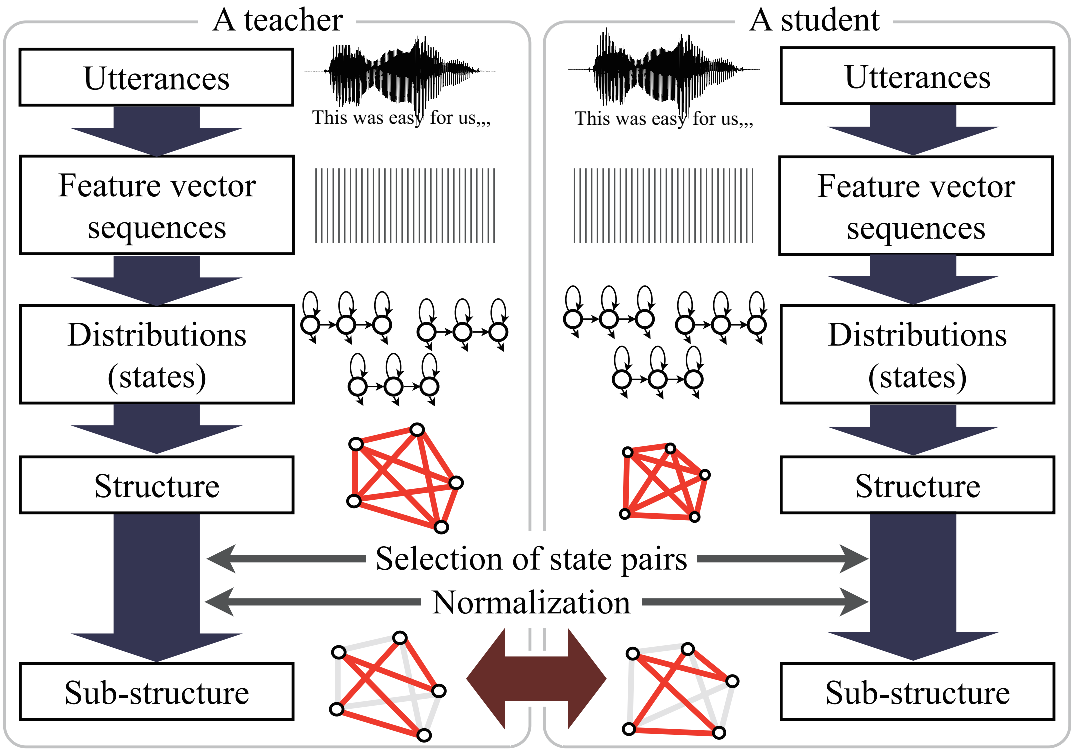


Figure 9.12: Procedure of extracting sub-structure

manual evaluation score vector is calculated and the result is shown by Fig.9.13. After that, sub-structure is built by deleting utterances from the structures together with the same cross validation. Then by comparing the new structural scores with the manual evaluation scores, the result in Fig.9.14 is obtained. In both the figures, X-axis means the number of added or deleted utterances and Y-axis means the correlation coefficient with the manual evaluated scores. And the line means the averaged correlation coefficient of all the speakers, the dot line means the correlation coefficient of two selected speakers.

About the averaged correlation coefficients in Fig.9.13, we can find that at the beginning, the result is a kind of over-fitting, then the result is raising and it begins to fall at last. The best steady result is about 0.3 when about 50 utterances are selected to build the structure. And about the selected speaker, the same trend can be found. In Fig.9.14, which is obtained when utterances was

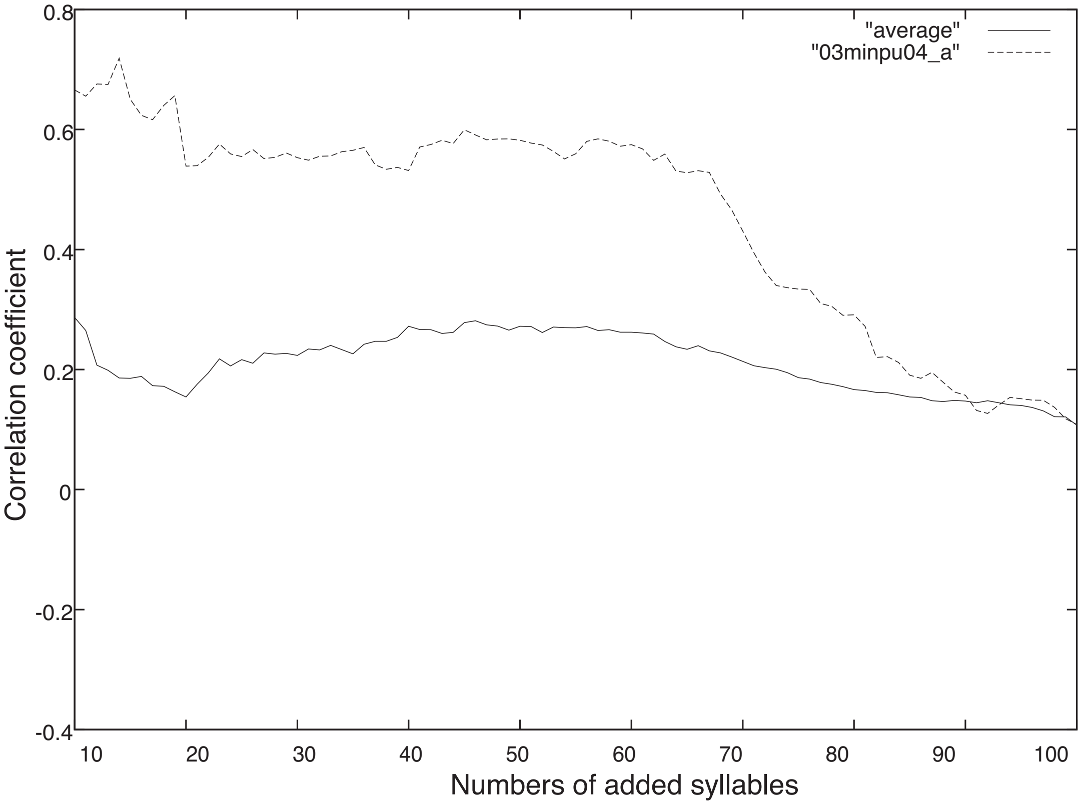


Figure 9.13: Building the sub-structures by adding utterances

deleted from the structures one by one, it is found that the averaged correlation coefficient is raising at the beginning and turns to over-fitting at the end. About the selected speaker, the same trend can also be found. Ant the best result is about 0.4 when 60 utterances are used to built the structures.

9.4 Conclusion

In this chapter, the structural method is applied to estimating the utterance similarity orders of dialects and pronunciation assessment of accented Mandarin. At the beginning, the dialect pronunciation structures can be built for dialect speakers. Then their dialect utterances can be compared with the pronunciation of standard Mandarin speakers using structural method and the utterance similar-

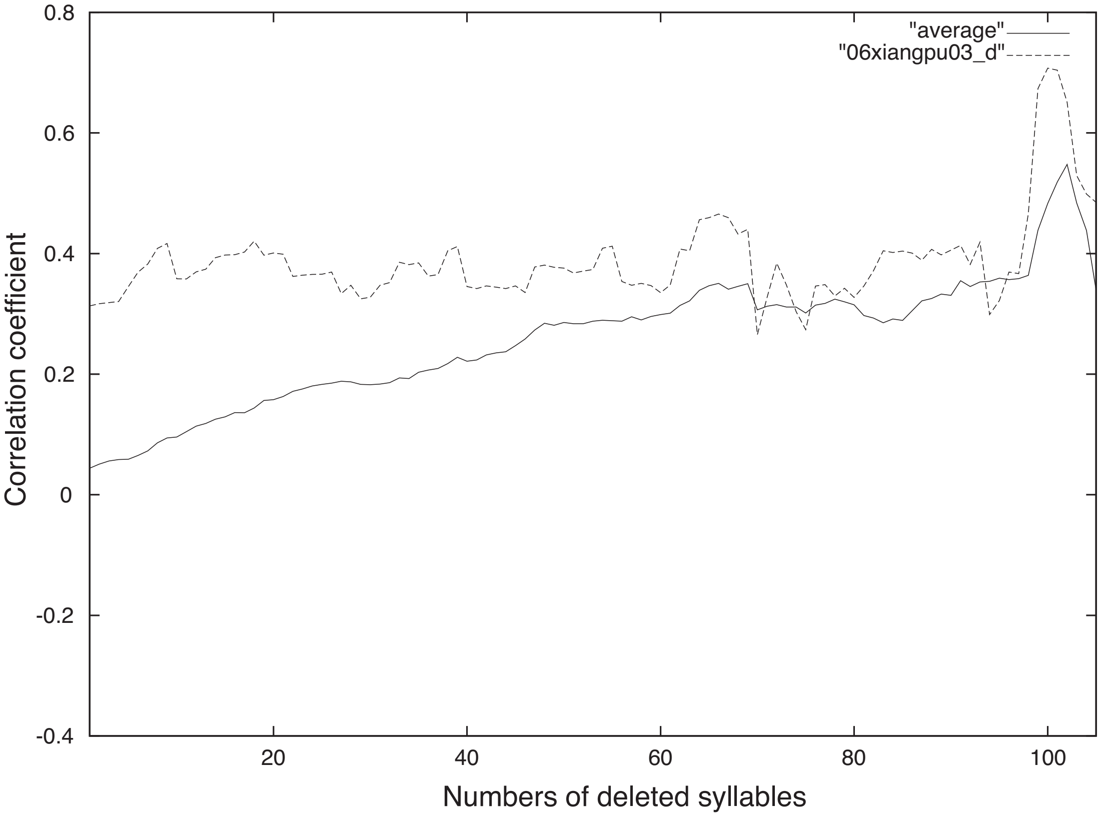


Figure 9.14: Building the sub-structures by deleting utterances

ity orders can be estimated. After that, experiments are also carried out using the dialect data of 2 Min speakers of different genders and the data of 2 standard Mandarin speakers of different genders. The results show that very similar similarity orders are obtained for the dialect speakers from the same region and the result is robust to the genders of the speakers. After that, the structural method is applied to accented Mandarin pronunciation assessment. At the beginning, the pronunciation structures of accented Mandarin are built and compared with the structures of standard Mandarin. Then a structural score is obtained for every utterance. After that, these utterances are evaluated manually and the manual evaluated sores are compared with the structural scores. Meanwhile, the structural scores are also compared with the result of recognized with a new built Mandarin recognizer. Then the correlation coefficients between these scores are calculated and the results are not very satisfactory. Considering the structures are built with some utterances that are very similar to each others, sub-structures are extracted to assess the accented Mandarin pronunciations. By adding or deleting utterances to built sub-structures, the pronunciations of accented Mandarin speakers are compared with the pronunciation of standard Mandarin speakers and the best correlation coefficient is obtained at about 0.4.

Chapter 10

Conclusions

10.1 Introduction

This thesis has investigated Chinese dialects analysis using structural pronunciation representation. In particular, pronunciation structure is proposed to represent Chinese dialects to extract the purely linguistic features and these features can be adopted to classify speakers based on their dialects and assess the pronunciations of different dialect or accented Mandarin speakers.

In modern speech technologies, speech is also represented by spectrum and which contains not only linguistic features but also extra-linguistic features like the age, gender, speaker, recording microphone and so on. However, for the problem of dialect-based speaker classification, only the linguistic features are needed and the extra-linguistic features should be removed. In order to solve this problem, speaker-independent acoustic models are trained using the data of many speakers and different normalization and speaker adaptation techniques are also needed in conventional speech recognition framework. And about the related studies of some linguists, in order to extract the linguistic and sociolinguistic differences between vowels of different speakers, different normalization techniques are always adopted. However, none of these methods can work well for the problem of Chinese dialect-based speaker classification. For this problem, the linguistic features invariant to extra-linguistic factors should be extracted from the dialect utterances of individual speakers.

Current situation of Chinese dialects are very complicated. There are hundreds kinds of dialects in China and they are mainly classified into several big dialect groups. Further, there are many different sub-dialects for every dialect group. Although these dialects are developed from the same root and inherited many common features, they are still quite different to each other grammatically, lexically, phonologically and phonetically. Sometimes, even for speakers from adjacent cities, their dialects are quite different and they may have difficulties in oral communication. Therefore, the method of building speaker-independent acoustic models using the data of different speakers can not work in our problem of Chinese dialect-based speaker classification, because it is very challenging to collect enough sub-dialect data to build dozens of sub-dialect acoustic models for one dialect region and the method of building sub-dialect models using the data of many speakers conflict with our goal of extracting the intra-dialect relations among speakers. Meanwhile, the linguistic method of vowel normalization also cannot work here because the information like the dialects of the speakers are needed for these normalization techniques but we don't know that before the dialect-based speaker classification experiment.

10.2 Summary of my works

In my works, pronunciation structure is proposed to represent Chinese dialect pronunciation and applied to dialect-based speaker classification and pronunciation assessment. Then several different classification experiments are carried out to prove that this method can extract the speaker-invariant purely linguistic features and classify speakers based on their dialects. After that, this approach is further proposed to be applied to calculating the utterance similarity orders between dialect pronunciations and pronunciation assessment of accented Mandarin.

In order to prove my proposal, dialect-based speaker classification experiment is carried out at the beginning. As the publicly available corpora cannot used for our problem, a new corpus of Chinese dialects is built with the dialect data of 17 dialect speakers. Then all the data are labeled manually and the final parts are converted into distributions. After that, for every speaker, the BDs between any pair of distributions are calculated and his/her pronunciation structure is built. By calculating the distances between the pronunciation structures of the speakers, they are classified based on their dialects and the result is independent to features of speakers like the gender and age.

After that, sub-dialect based speaker classification experiment is carried out. A new corpus of sub-dialects is built with 16 speakers from 4 sub-dialects of Mandarin and the data is analyzed under the same acoustic condition like last experiment. Then sub-dialect pronunciation structures are built and these speakers are classified by calculating the distances between their pronunciation structures. The result shows the speakers are mainly classified by their sub-dialects with one exception that 4 speakers from one same sub-dialect region are classified to different sub-trees. About the reasons for it, it is considered that these speakers are from different sub-sub-dialect regions and their sub-dialect may be affected by other sub-dialects that they are living. Meanwhile, the sub-dialect regions of these speakers are obtained by traditional linguistic classification, which is not based on the acoustic features of dialect pronunciations like our structural method.

Several comparison experiments are designed to prove that our method can extract the speaker-invariant linguistic features from the pronunciations of Chinese dialect speakers, no matter which kind of dialects are they speaking. For the new comparison experiments, a new corpus is built with the speakers from 10 sub-dialects of 5 dialect regions. Then corresponding to this corpus, every utterance is mimicked linguistically by an expert of Chinese dialects and a new corpus with minimum speaker differences is built. After that, using the original and mimicked data, dialect-based speaker classification experiments are carried out. We find the two results are almost the same as each other, although one is obtained using the dialect data spoken by different speakers and the other is obtained using the dialect data with fixed speaker identity. So it means our method is really invariant to speakers.

Also, our method of structural pronunciation comparison is compared with conventional spectral comparison. At the beginning, corresponding to the original and mimicked dialect data used above, new data are converted just like they are pronounced by a very tall speaker and very short speaker and data sets with maximum speaker differences are built. Then using these data, classification experiments based on spectral comparisons are carried out. The results show that the classifications are affected greatly by the speaker features. After that, they are classified using our structural method and the results show that these speakers are well classified by their dialects and it is not affected by the speaker differences. So our method is proved again that it can classify speakers based on their dialects by extracting the purely linguistic features and unlike the conventional spectral comparison, the result is not affected by feature of speakers.

At last, the structural method is applied to compare dialect pronunciations and estimate the utterance similarity orders between any two speakers. Using the dialect data of 2 Min speakers of different genders and the data of 2 standard Mandarin speakers of different genders, some experiments are carried out to estimate the utterance similarity orders among them. The results show that very similar similarity orders are obtained for the dialect speakers from the same region and the result is robust to the genders of the speakers. Also, this structural method is applied to pronunciation assessment of accented Mandarin. At the beginning, the pronunciation structures of accented Mandarin are built and compared with the structures of standard Mandarin. Then a structural score is obtained for every utterance. After that, these utterances are evaluated manually and the manual evaluated sores are compared with the structural scores. Meanwhile, the structural scores are also compared with the recognition score obtained by a new built Mandarin recognizer. However, the correlation coefficients between these scores are not satisfactory, although some correlations can be found by the results. So considering the reasons for it, sub-structures are built to assess the accented Mandarin pronunciations. By adding or deleting utterances to built sub-structures, the pronunciations of accented Mandarin speakers are compared with standard Mandarin speakers and the best correlation coefficient is obtained at about 0.4.

10.3 Future works

As we already proved that the structural pronunciation representation can extract the purely speaker features from Chinese dialects and speaker-invariant dialect-based classification can be achieved, this approach can be applied to many future works. For example, a dialect recognition system can be built by building the dialect pronunciation structures for all the major dialect or sub-dialect and the results can be further applied to speech recognition of different dialects. Also, for the linguists, this method can be used to calculate the acoustic distances between any two dialects and a new atlas of Chinese dialect based on the acoustic features of dialects can be built.

Furthermore, the structure method is already applied to estimating the utterance similarity orders between any two dialect speakers and pronunciation assessment of accented Mandarin. About the estimation of the utterance similarity orders, as the speaker-invariance of this method is already proved, it can be applied to find the relationship between the individual utterance of different dialects. About the experiment of pronunciation assessment of accented Mandarin, only using the data of two standard Mandarin speakers to calculate the structural scores and using 3 levels to evaluate the pronunciation manually, the correlation coefficient between them is got at about 0.4. I am thinking that if more data of standard Mandarin are obtained and the manual evaluation is more accurately, higher correlation coefficient between these two scores would be obtained.

References

[1] Cynthia G. Clopper and David B. Pisoni, “Some acoustic cues for the perceptual categorization of American English regional dialects,” Journal of Phonetics, 32:1, pp.111–140 (2000) 1

[2] Labov, W. and Ash, S. and Boberg, C., “The Atlas of North American English: Phonology, Phonetics, and Sound Change: a multimedia reference tool,” Walter De Gruyter Inc (2006) 1, 23

[3] Helen. Chen, “Calculation of phonological similarity between dialects,” Language Sciences, 5:1, pp.23–31 (2006) 1

[4] Nerbonne, J. and Heeringa, W. and van den Hout, E. and van der Kooi, P. and Otten, S. and Van De Vis, W., “Phonetic Distance between Dutch Dialects,” Proceedings of CLIN’95, Antwerp, pp.185–202 (1996) 1

[5] Tsai, W.H. and Chang, W.W., “Discriminative training of Gaussian mixture bigram models with application to Chinese dialect identification,” Speech Communication, 36:3–4, pp.317–326 (2002) 1

[6] Purnell, T. and Idsardi, W. and Baugh, J., “Perceptual and phonetic experiments on American English dialect identification,” Journal of Language and Social Psychology, 18:1, pp.10–30 (1999) 1

[7] Shen, W. and Chen, N. and Reynolds, D., “Dialect recognition using adapted phonetic models,” Proc. InterSpeech, pp.763–766 (2008) 1, 18

[8] Lee, K.F. and Hon, H.W., “Speaker-independent phone recognition using hidden Markov models,” IEEE Transactions on Acoustics, Speech and Signal Processing, 37:11, pp.1641–1648 (1989) 1

[9] Byrne, W., Beyerlein, P., Huerta, J., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D. and Wang, W., “Towards language independent acoustic modeling,” IEEE International Conference on Acoustics Speech and Signal Processing, 2 (2000) 2

[10] Hindle, D.. 1978. Approaches to vowel normalization in the study of natural speech. In D. Sankoff (ed.), Linguistic Variation: Models 2

[11] Disner, S.F., 1980. Evaluation of vowel normalization procedures. Journal of the Acoustical Society of America 67:253-61 2

[12] Adank, P., Smits, R. and Van Hout, R., 2004. A comparison of vowel normalization procedures for language variation research. Journal of the Acoustical Society of America 116:3099-107. 2

[13] Haugen, E., “Dialect, language, nation,” American Anthropologist, 68:4, pp.922–935 (1966) 2

[14] Norman, J., “Chinese,” Cambridge Univ Pr, (1988) 2, 6, 8

[15] Li, C.N. and Thompson, S.A., “Mandarin Chinese: A functional reference grammar,” Univ of California Pr (1989) 2, 6

[16] Thurgood, G. and LaPolla, R.J., “The Sino-Tibetan languages,” London : Routledge (2003) 2, 6, 8

[17] Weinreich, M., “Der Yivo un di Problemen fun Undzer Tsayt, Yivo-Bleter”, 25(1): 13 (1945) 2, 8

[18] Campbell, J., “Chinese language FAQ. Chinese language FAQ. Glossika Language Web,” http://www.glossika.com/en/dict/faq.php (2004) 2, 8

[19] Yuan Jiahua, “HanYu FangYan GaiYao,” Language & Culture Press (2000) 2, 8, 10, 30, 43

[20] Hou Jingyi, “XianDai HanYu FangYan GaiLun,” ShangHai Education Publishing House (2002) 2, 8, 43

[21] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889–892 (2005) 3, 30

[22] N. Minematsu, "Theorem of the invariant structure and its derivation of speech gestalt," Proc. Workshop on Speech Recognition and Intrinsic Variations, pp.47–52 (2006) 3

[23] S. Asakawa, N. Minematsu and K. Hirose, "Multi-stream parameterization for structural speech recognition," Proc. ICASSP, pp.4097–4100 (2008) 3

[24] Y. Qiao and N. Minematsu, "f-divergence is a generalized invariant measure between distributions," Proc. INTERSPEECH, pp.1349–1352 (2008) 3, 28

[25] D. Saito, S. Asakawa, N. Minematsu and K. Hirose, "Structure to speech – speech generation based on infantlike vocal imitation –," Proc. INTERSPEECH, pp.1837–1840 (2008) 3

[26] N. Minematsu, K. Kamata, S. Asakawa, T. Makino, T. Nishimura and K. Hirose, "Structural assessment of language learners' pronunciation," Proc. INTERSPEECH, pp.126–129 (2007) 3

[27] Chen, M.Y., "Tone sandhi: Patterns across Chinese dialects," Cambridge Univ Pr (2000) 8

[28] Chinese Academy of Social Sciences, "Language Atlas of China," Hong Kong: Longman Group (1988) 8, 44, 53

[29] Duanmu, S., "The Phonology of Standard Chinese," Oxford University Press Oxford (2000) 8

[30] Chen, T., Huang, C., Chang, E. and Wang, J., "Automatic accent identification using Gaussian mixture models," IEEE Workshop on Automatic Speech Recognition and Understanding, pp.343–346 (2001) 10

[31] Gold, B. and Morgan, N., "Speech & Audio Signal Processing," Wiley-India (2001) 15

[32] Huang, X., Acero, A., Hon, H.W., "Spoken language processing," Prentice Hall PTR (2001) 15

[33] Schafer, RW and Rabiner, LR, "Digital representations of speech signals," Proceedings of the IEEE, 63:4, pp.662–677 (1975) 15

[38] Stevens SS, Volkmann, J and Newman, EB, "A scale for the measurement of the psychological magnitude pitch," The Journal of the Acoustical Society of America, 8, pp.185–190 (1937) 16

[37] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, 28:4, pp.357–366 (1980) 16

[36] H. Liao and M. Gales, "Uncertainty decoding for noise robust speech recognition," Doctor thesis, University of Cambridge (2004) 16

[37] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, 28:4, pp.357–366 (1980) 16

[38] Stevens SS, Volkmann, J and Newman, EB, "A scale for the measurement of the psychological magnitude pitch," The Journal of the Acoustical Society of America, 8, pp.185–190 (1937) 16

[39] Rabiner, LR, "A tutorial on hidden Markov models and selected applications inspeech recognition," Proceedings of the IEEE, 77:2, pp.257–286 (1989) 17

[40] Hwang, M.Y., Wang, W., Lei, X., Zheng, J., Cetin, O. and Peng, G., "Advances in Mandarin broadcast speech recognition," Proc. Interspeech, pp.2613–2616 (2007) 17

[41] Droppo, J. and Acero, A., "Noise robust speech recognition with a switching linear dynamic model," Proc. ICASSP, pp.953–956 (2004) 17

[42] Schultz, T. and Waibel, A., "Language-independent and language-adaptive acoustic modeling for speech recognition," Speech Communication, 35:1, pp.31–52 (2001) 18

[43] Schultz, T. and Kirchhoff, K., "Multilingual speech processing," Academic Press (2006) 18

[44] Zissman, M.A. and Berkling, K.M., "Automatic language identification," Speech Communication, 35:1–2, pp.115–124 (2001) 18

[45] Ma, B. and Zhu, D. and Tong, R., "Chinese dialect identification using tone features based on pitch flux," Proc. ICASSP, 1, pp.1029–1032 (2006) 18

[46] Yanguas, L.R., O'Leary, G.C. and Zissman, M.A., "Incorporating linguistic knowledge into automatic dialect identification of Spanish," Fifth International Conference on Spoken Language Processing (1998) 18

[47] Beattie, V.L., "Multi-dialect speech recognition method and apparatus," Acoustical Society of America Journal, 107, pp.1817 (2000) 18

[48] Liu, F.H., Stern, R.M., Huang, X. and Acero, A., "Efficient cepstral normalization for robust speech recognition," Proceedings of ARPA Speech and Natural Language Workshop, pp.69–74 (1993) 20

[49] Zhan, P. and Waibel, A., "Vocal tract length normalization for large vocabulary continuous speech recognition," Proc. EUROSPEECH, pp.2527–2530 (1997) 20

[50] Leggetter, CJ and Woodland, PC, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer speech and language, 9:2, pp.171–185 (1995) 20

[51] Zhang, B. and Matsoukas, S., "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," International Conference on Acoustics, Speech, and Signal Processing, pp.925–928 (2005) 20

[52] http:ncslaap.lib.ncsu.edutoolsnormabout_normalization1.php 21

[53] Miller, J.D., “Auditory-perceptual interpretation of the vowel,” The journal of the Acoustical society of America, 85, pp.2114 (1989) 22

[54] Adank, P., Smits, R. and Van Hout, R., “A comparison of vowel normalization procedures for language variation research,” The Journal of the Acoustical Society of America, 116, pp.3099 (2004) 23

[55] Lobanov, BM., “Classification of Russian vowels spoken by different listeners,” Journal of the Acoustical Society of America, 49, pp.606–608 (1971) 23

[56] Nearey, T.M., “Phonetic feature systems for vowels,” Indiana University Linguistics Club (1978) 23

[57] Feng, S., “The Vowel pattern of Beijing Mandarin,” Nankai Linguistics (2002) 23

[58] V.S. Richard, “Handbook for Lexicon Based Dialect Fieldwork,” Zhonghua Book Company (2006) 24, 30

[59] Institute of Linguistics of CASS, “Hanyu DiaoCha ZiBiao,” The Commercial Press (2007) 24, 30, 49, 56

[60] C.C. Cheng, “Syllable-based dialect classification and mutual intelligibility,” Chinese Languages and Linguistics I Chinese Dialects, pp.145–177 (1992) 24

[61] Pitz, M. and Ney, H., “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech and Audio Processing, 13, pp.930–944 (2005) 27

[62] D. Saito, N. Minematsu, and K. Hirose, “Decomposition of rotational distortion caused by VTL difference using eigenvalues of its transofmation matrix,” Proc. INTERSPEECH, pp.1361-1364 (2008) 30, 56

[63] Willett, P., “Recent trends in hierarchic document clustering: a critical review,” Information Processing & Management, 24:5, pp.577–597 (1988) 36

[64] Kawahara, H., Masuda-Katsuse, I. and de Cheveign, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187-207 (1999) 57

[65] M. Suzuki, N. Minematsu, D. Luo and K. Hirose, “Sub-structure-based estimation of pronunciation proficiency and classification of learners,” Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU’2009), pp.574-579 (2009) 81

Published papers

Journal paper

[1] X. Ma, N. Minematsu, Y. Qiao, R. Xu, A. Li and K. Hirose, “Speaker classification using speaker invariant dialectal features extracted through pronunciation structure,” Speech Communication, vol.X, no.Y, pp.xx-yy (2010, submitted)

Peer-reviewed conference paper

[2] N. Minematsu, M. Takazawa and X. Ma, “Pronunciation clinic & dialect-based speaker classification,” Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP’2009, Show & Tell), (2009)

[3] X. Ma, N. Minematsu, A. Nemoto, M. Takazawa, Y. Qiao and K. Hirose, “Dialect-based speaker classification of Chinese using structural representation of pronunciation,” Proc. Speech and Computer (SPECOM), pp.350–355 (2009)

[4] X. Ma, N. Minematsu, A. Nemoto, M. Takazawa, Y. Qiao and K. Hirose, “Structural analysis of Chinese dialect speakers and their automatic classification,” Proc. National Conference on Man-Machine Speech Communication, pp.440–445 (2009)

[5] X. Ma, A. Nemoto, N. Minematsu, Y. Qiao and K. Hirose, “Structural analysis of dialects, sub-dialects, and sub-sub-dialects of Chinese,” Proc. INTERSPEECH, pp.2219–2222 (2009)

[6] X. Ma, R. Xu, N. Minematsu, Y. Qiao, K. Hirose and A. Li , “Dialect-based speaker classification using speaker-invariant dialect features,” Proc. International Symposium on Chinese Spoken Language Processing, pp.xx-yy (2010, submitted)

Non peer-reviewed conference papers

[7] X. Ma, M. Takazawa, N. Minematsu and K. Hirose , “Chinese dialect classification using acoustic universal structure in speech,” Proc. Autumn Meeting of Acoust. Soc. Japan, pp.405–408 (2008)

[8] X. Ma, N. Minematsu, Y. Qiao, K. Hirose, A. Nemoto and F. Shi , “Dialect-based speaker classification of Chinese using acoustic features invariant with extra-linguistic factors,” IEICE Technical Report, SP2008-109, pp.179–184 (2008)

[9] X. Ma, A. Nemoto, N. Minematsu and F. Shi , “Development of a Chinese speech corpus covering inter-dialect phonological differences,” Proc. Spring Meeting of Acoust. Soc. Japan, pp.179–184 (2009)

[10] X. Ma, N. Minematsu, A. Nemoto, Y. Qiao and K. Hirose, “Structural analysis of Chinese dialects and its experimental application to pronunciation assessment,” IEICE Technical Report, SP2009-45, pp.25–30 (2009)

[11] 峯松信明, 鈴木雅之, 高澤真章, 馬学彬, 中村綾乃 , “発音クリニック　〜音声の構造的表象を用いた外国語・方言発音分析〜,” 情報処理学会講演論文集, 5, pp.303–304 (2010)

[12] X. Ma, N. Minematsu, R. Xu, A. Li, K. Hirose, “Speaker-invariance verification of dialect pronunciation structure applied in dialect-based speaker classification,” Proc. Autumn Meeting of Acoustic. Soc. Japan, pp.xx-yy (2010, submitted)