社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

中国語方言の構造分析とその発音評価に向けた実験的検討

馬 学彬† 峯松 信明† 根本 晃†† 喬 宇† 広瀬 啓吉†

† 東京大学 〒 113–8656 東京都文京区本郷 7–3–1
†† 南開大学 〒 300071 天津南開区衛津路 94 号 (中国)
E-mail: †{xuebin,mine,qiao,hirose}@gavo.t.u-tokyo.ac.jp, ††akiranmt@hotmail.com

あらまし　中国語方言は、地理的社会的言語的要因により非常に複雑な状況にある。中国には数百種類もの方言、下位方言、更なる下位方言があり、方言話者が標準語を話す場合にはそれらの影響を強く受ける。通常、音声の音響的特徴はスペクトルによって表されるが、これには方言的特徴の他に、性差、年齢差、個人差などの非言語的特徴も含まれているため、中国語方言話者に対するの自動発音評価は容易ではない。本研究では、音声の構造的表象を用いて非言語的情報を効果的に除去した後になまりのある発音を評価する手法に関する実験的検討を行なう。この手法では、まず漢字音セットに対する方言の録音を行った後、話者ごとに発音構造を構築する。そして他の話者との構造間距離を比較することで、発音を評価する。本稿では同じ漢字の方言発音と標準語の発音を比べることで、発音の類似度を算出し、それが性別に対して高い頑健性をもつことを示した。

キーワード　中国語方言、非言語的特徴、 変換の不変性、構造的分析、発音評定

Structural analysis of Chinese dialects and its experimental application to pronunciation assessment

Xuebin MA†, Nobuaki MINEMATSU†, Akira NEMOTO††, Yu QIAO†, and Keikichi HIROSE†

† The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo 113–8656 Japan
†† Nankai University, 94 Wenjin Rd, Nankai, Tianjin, 300071 P.R. China
E-mail: †{xuebin,mine,qiao,hirose}@gavo.t.u-tokyo.ac.jp, ††akiranmt@hotmail.com

Abstract　The current situation of Chinese dialects is very complicated, not only because there are hundred kinds of dialects, sub-dialects and sub-sub-dialects, but also because many dialect speakers are speaking Mandarin with different regional accents affected by their dialects. If dialect utterances are represented acoustically, their acoustic features depend both on dialectal linguistic features and extra-linguistic features such as the age, gender, microphone, etc. Then, the automatic pronunciation assessment of Chinese dialects becomes a challenging task. In this paper, we propose to analyze the dialect pronunciation by the structural representation, which was originally used to remove extra-linguistic information from speech. After collecting the utterances of some selected Chinese written characters, the pronunciation structures are built for dialect speakers and their pronunciations are compared by the distances among the structures. Then with the dialect utterances of 16 native dialect speakers, some experiments are carried out to analyze their pronunciations using the structures and linguistically-reasonable results are obtained. At last, by comparing the pronunciation structures of dialect speakers with those of two standard Mandarin speakers, the pronunciation similarity orders of the characters are obtained. Through these experiments, our approach is proved again that it works well in extracting the linguistic features from dialect utterances and shows high independence of the extra-linguistic features.

Key words　Chinese dialects, extra-linguistic features, invariance, structural analysis, pronunciation assessment

1. Introduction

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. But in the case of pronunciation assessment of dialect speakers, we should only focus on the acoustic features which are relevant to dialectal information and irrelevant to extra-linguistic information. It is because the acoustic differences between two utterances of the same linguistic content spoken by a very tall adult and a very short child are sometimes larger than the acoustic differences between a Mandarin utterance of an adult and its dialectal version of that adult. For this problem, in automatic dialect speech recognition systems, speaker-independent acoustic models are often built for each dialect by collecting utterances from thousands of different speakers of that dialect but speaker adaptation or normalization techniques are still required. Nevertheless, this approach cannot be accepted in pronunciation assessment of dialect speakers. Strictly speaking, speakers of the same dialect are often speakers of different sub-dialects and their acoustic features relevant only to dialectal information are needed and, For these reasons, dialect pronunciation assessment cannot be attained by training individual dialect models with utterances from many speakers.

In our previous study, in order to capture the purely linguistic information and remove the variations caused by extra-linguistic features, a structural representation of speech was proposed [1], [2]. After modeling the speech variations caused by extra-linguistic factors mathematically, this speech structure is calculated by extracting speaker-invariant speech contrasts or dynamics and shows high speaker independence. Currently, this speech structure was already applied to speaker independent ASR system [3], [4]. Structure-based ASR showed much higher performance than widely used methods especially in mismatched conditions, although it was realized only with a small number of speakers and without explicit speaker adaptation or normalization. Further, the structure was also applied for helping Japanese learning English [5], speech synthesis [6], dialect-based speaker classification [7] with satisfactory results obtained.

In this paper, the speech structure is applied to pronunciation assessment of different dialect speakers of Chinese. As this structure can extract the acoustic features relevant only to dialectal differences and is invariant with extra-linguistic factors, pronunciation of different dialect speakers can be assessed purely on their linguistic features. At the beginning, some fundamental knowledge about Chinese dialects and the current complicated situations are described in Section 2. Then in Section 3, how to model the extra-linguistic features and build the dialectal speech structures is presented. In Section 4, building the comparable structures among different dialects is shown according to linguistic studies. In Section 5, with the dialect speech of 16 speakers from 5 dialect regions, experiments of structural assessment of their pronunciation are carried out and discussed. Further, the estimation of the vowel similarity between pronunciation of dialects and that of Mandarin is investigated. All the results show high validity and accordance to linguistic studies. At last, the paper is concluded in Section 6.

2. Fundamentals of Chinese dialects

In China, there are hundred kinds of dialects. Although they can be classified into different groups according to different criterions, it is widely accepted by Chinese traditional dialectology that they can be grouped into 7 big dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue, Min) [8]. Further, most of these dialect regions can be further grouped into many different sub-dialects and sub-sub-dialects. For example, the dialects of GuanHua (also called Mandarin) dialect region can be grouped into 8 sub-dialects and many sub-sub-dialects [9]. In fact, all the dialects and sub-dialects have been developed from Old Chinese and Middle Chinese, and they have inherited a lot of common features. They are sharing the same written characters, similar sound systems, the same phonological structure and similar phonetic features, etc. For instance, every Chinese character is pronounced as mono-syllable with the same structure: a combination of an initial at the beginning, a final at the end and a tone. However, there are still many differences among these dialects grammatically, lexically, phonologically and phonetically in varying degrees and peoples from different dialect regions sometimes have difficulty in oral communication. Especially, the dialects of different big dialect regions are always regarded as different languages just like French and Italian.

Since 1956, standard Mandarin has been popularized all over the country as the official language. Then, many dialect speakers began to learn Mandarin just like learning a second language. However, affected by their native dialects, many of them speak Mandarin with some regional accents in different degrees. Sometimes, one can guess their native dialects easily by hearing their accented Mandarin if he/she has some knowledge of these dialects. On the other hand, as standard Mandarin becomes more and more popular and many people of different dialect regions are moving all over the country, some dialects are affected and losing some of their own dialectal features. However, these dialects, especially the major dialects, are still widely used and will continue to be used in the future.

In brief, the current situation of Chinese dialects is very complicated and it is really a challenge to assess the pronunciation of dialect speakers using a computer automatically. It is not only because the linguistic features of dialectal pro-

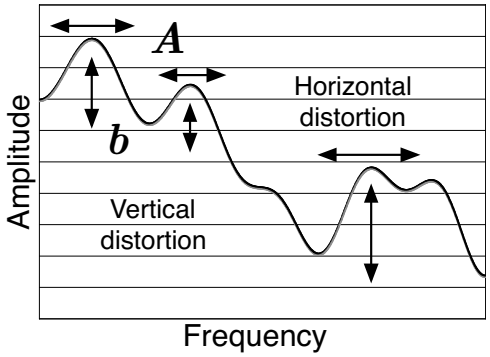


Fig. 1 Spectral distortions caused by matix A and vector b

nunciations are always affected by extra-linguistic features acoustically, but also because every speaker has his/her own dialect. Strictly speaking, the pronunciations of two speakers of the same dialect show somewhat different linguistic features. So in order to assess the pronunciation of different speakers, it is necessary to focus on the dialectal features of every speaker that are invariant with extra-linguistic factors.

3. Pronunciation structure of dialects

3.1 Modeling the extra-linguistic features

When speech is represented acoustically by spectrum, the inevitable extra-linguistic factors can be classified as either of two kinds of distortions according to their spectral behaviors: convolutional and linear transformational distortions. Convolutional distortions are caused by extra-linguistic factors such as different recording microphones and vocal tract length differences are the typical reason of linear transformational distortions [10]. If a speech event is represented by cepstrum vector c, the convolutional distortion is 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 spectral distortions caused by inevitable extra-linguistic features can be modeled by $c' = Ac + b$, known as affine transformation. The distortion is schematized by Fig. 1. The horizontal and vertical distortions correspond to the distortions due to matrix A and vector b, respectively.

3.2 Speaker-invariant structural representation

As the acoustic features caused by extra-linguistic factors can be modeled as affine transformations, we can use affine-invariant features to obtain speech features invariant to extra-linguistic variations. In fact, every speech event can be captured as a distribution and event-to-event distances are calculated as Bhattacharyya Distance (BD).

$$BD(p_1, p_2) = -\ln \oint \sqrt{p_1(c)p_2(c)}dc, \qquad (1)$$

Then a distance matrix can be obtained by calculating BDs between any pair of speech events. Since BD is invariant

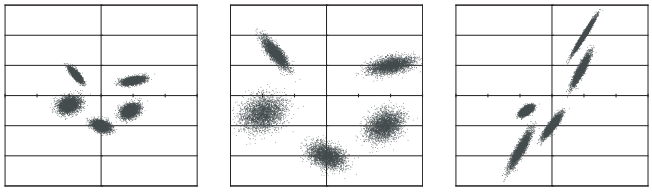


Fig. 2 The invariant underlying structure among three data sets

with affine transformation, the built structure is invariant to extra-linguistic factors. Fig.2 shows an example of the invariant underlying structure among three sets of speech events. Although these sets of events look very different to each other in the Figure, any set of them is obtained as affine transform of either of the other two sets and their BD-based distance matrices are identical. Because a distance matrix can represent uniquely its geometrical shape composed of all the speech events, we call this matrix a pronunciation structure of these speech events. Then with the utterances of dialect speakers, we can build structural representations of dialect pronunciation which are sensitive to dialectal features but invariant to extra-linguistic factors. In other words, if the structures are built separately from two speakers of the same dialect, structural difference between them is small. If they are built from two speakers of different dialects, the difference will be large but it is independent of age and gender.

4. Comparable pronunciation structures

In order to assess the pronunciation of different dialect speakers using structural representation, comparable dialect structures have to be built using their dialectal utterances of the same set of some 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 inventory of these phonological units changes from dialect to dialect. As 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 and dialect pronunciation structures for the individual dialects can be calculated using the utterances of whole syllables or only their final parts. If we can select a set of characters covering the most of the phonological units in the dialects, comparable structures can be built.

In these years, many Chinese linguists are studying Chinese dialects and some of them are focusing on the phonological features and the relationships among the dialects. By checking the historical changes in the pronunciation of some written characters and their current pronunciation in different dialects, the phonological differences among dialects can

Table 1 Detailed information of dialect speakers

ID	Dialect	Hometown	Sub-dialect	Gender
01	Hakka	TongGu	TongGu	F
02	Hakka	TongGu	TongGu	F
03	Gan	FuZhou	YiLiu	F
04	Gan	ShangGao	FuGuang	F
05	Xiang	Xiangtan	ChangYi	F
06	Xiang	Xiangtan	ChangYi	F
07	Xiang	JiShou	JiShu	F
08	Xiang	ShaoYang	LouShao	F
09	Yue	GuangZhou	GuangFu	F
10	Yue	FoShan	GuangFu	M
11	Yue	FoShan	GuangFu	F
12	Yue	GuangZhou	GuangFu	F
13	Min	XiaMen	QuanZhang	F
14	Min	XiaMen	QuanZhang	M
15	Min	JiJang	QuanZhang	F
16	Min	QuanZhou	QuanZhang	F

be compared. For example, in [11], the historical pronunciations and modern dialectal pronunciations of the commonly used written characters are all listed. Then based on these studies, some specific lists of written characters are often adopted by linguists to check the features of corresponding initials, finals and tones in different dialects [12] [13]. In [13], which is written by linguists in the Institute of Linguistics of Chinese Academy of Social Sciences, three different lists of written characters are shown for checking the dialectal features of tones, initials and finals, separately. Then using the dialectal utterances of these characters, the speaker-invariant but dialect-sensitive pronunciation structure can be built for every speaker. 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.

5. Analysis of dialect pronunciations

5.1 Speech material of dialects

In order to assess the pronunciation of speakers from different dialect regions, the written characters covering the differences among these dialects should be selected to build the pronunciation structure for every speaker. Here, a list of written characters in [13], which is used for checking the dialectal finals by linguists, is adopted as the reading materials for dialect speakers.

Table 2 Selected characters for dialect pronunciation assessment

Characters	辣、架、蛇、落、野、 月、資、知、耳、第、虛、 木、北、桂、桃、藥、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/

Table 3 Acoustic analysis condition

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

The subjects are 16 speakers from 11 cities belonging to 5 big dialect regions. They are all undergraduate students of Nankai University and have no background of other languages before entering the university in Tianjin. They were selected after their language backgrounds were checked to ensure they were brought up in the same dialect regions and their parents are also the native speakers of that dialect. More information such as the hometown, the sub-dialect region, gender of these speakers can be found in Table 1, where every speaker is given an ID number which is used in the following experiments.

All the recordings were carried out in quiet rooms with a supervisor, so the data are all expected to be clean. Before the recording, the dialectal pronunciations of all the reading characters were checked by every speaker. Then the recording was carried out with a 48KHz linear PCM recorder of Sony PCM-D1. Every speaker was asked to read the selected characters in their native dialects three times. Then all the data were labeled phonetically and manually by linguistic students. During this step, the utterances of some characters were discarded as their pronunciations were different obviously to the speakers from the same sub-dialect region. Then after checking the spectrum and raw files, every syllable was labeled into two parts, initial and final, with transcriptions mainly developed from Chinese Pinyin. Then the final part of every syllable is modeled as a single Gaussian distribution under the acoustic conditions shown in Table 3.

5.2 Structural analysis of dialect pronunciation

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. 3. Then the minimum sum of the distances between the corresponding two points of the two structures can be obtained with the best overlap. In [1], it was experimentally proved that

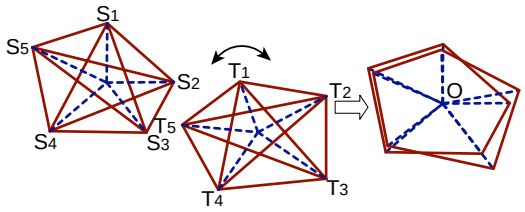


Fig. 3 Distance calculation after shift and rotation

the minimum sum can be approximately calculated as Euclidean distance between two distance matrices. Following is the detailed computing formula:

$$D_1(A,B)=\sqrt{\frac{1}{M}\sum_{i<j}(A_{ij}-B_{ij})^2}, \qquad (2)$$

where A_{ij} and B_{ij} mean the (i,j) element of matrices A and B, respectively. M means the number of the syllables.

After the pronunciation structures of dialect speakers are built using their dialectal utterances, their pronunciation can be assessed by the distances D_1 among the speakers. Then the result can be shown by a clustering tree built with these distances. It is supposed that the pronunciation of the speakers from the same dialect or sub-dialect regions, i.e. similar dialect pronunciations, will be clustered near to each other in the result. Here, we adopted Ward's bottom-up clustering method and the obtained result is shown by Fig. 4, where the pronunciation structure of every speaker is represented by the speaker ID in Table 1 and different colors show different dialect regions.

In this figure, we can see that the speakers from the same dialect regions are all clustered together. Further, the speakers from the same cities and sub-dialect regions are also clustered near to each other. For example, in the case of speaker 01 and 02 or speakers 13 and 14, who are from the same cities and supposed to have very similar dialect pronunciation, their structures are exactly clustered near to each other in the result. Besides, we can also find that the pronunciation structures of speakers from Hakka, Gan and Xiang are clustered into a big sub-tree while the speakers from Yue and Min are clustered into another big sub-tree. One reason is that speakers 01 and 02 are from a sub-dialect of Hakka which is located at the middle of Gan dialect region and the speakers from Xiang dialect region are also near to Gan dialect region geographically [14]. Meanwhile, it is claimed by some linguists that these three dialects are very near to each other genetically, geographically, phonologically and are more similar to each other than Yue and Min dialects [9].

5.3 Estimation of vowel similarity to Mandarin

The above result is obtained by comparing the whole pronunciation structures of different speakers and this procedure can be decomposed into the comparison of individual vowels:

$$d(A,B,v)=\sum_{v}|A_{vi}-B_{vi}|, \qquad (3)$$

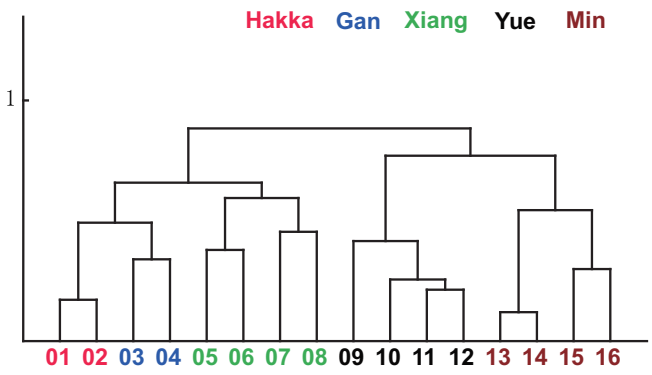
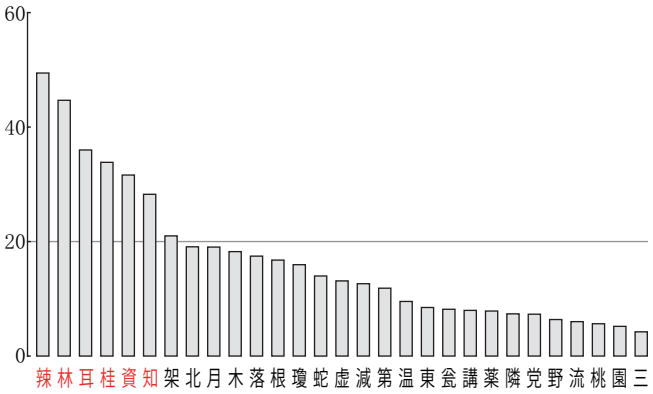


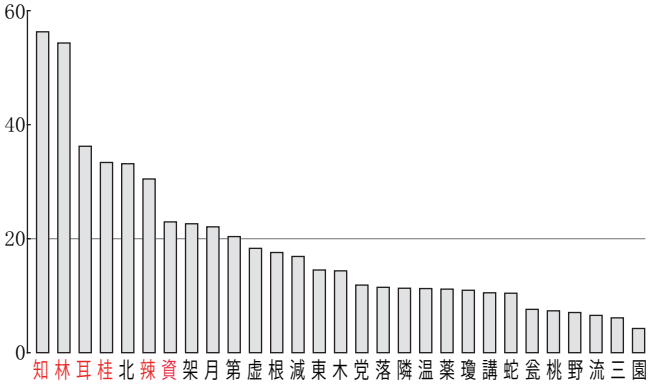
Fig. 4 Clustering of different dialect pronunciations

where A and B mean the matrices of two compared pronunciation structures, v means the vowel 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 vowel of the largest d means the pronunciation of this vowel is most dissimilar to the Mandarin pronunciation of speaker B. And the vowels with high values of d also mean that more attention should be paid to the vowels when the dialect speaker is learning Mandarin.

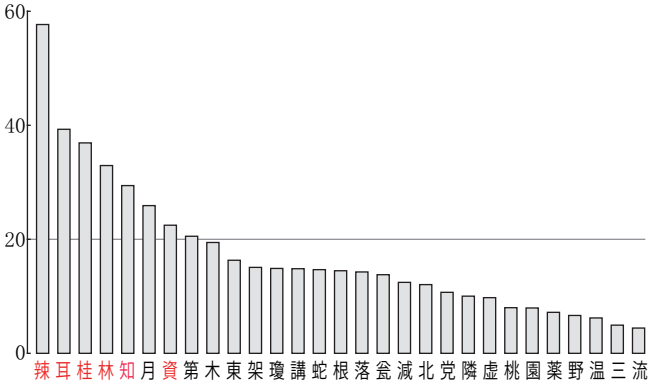
For this experiment, two standard Mandarin speakers, one male and one female, joined our recording. Using their utterances of the selected characters in Table 2, their pronunciation structures were built by the same approach. Then by checking each vowel using *d(A,B,v)*, the vowel similarity orders between this dialect speaker and a Mandarin speaker can be obtained. Here, in order to testify this approach, especially as to its independence of extra-linguistic features, Min speakers 13 and 14 (13 is a female and 14 is a male) are selected and the final parts of their utterances are used to calculate the similarity orders to the pronunciation of two Mandarin speakers. The results are shown by Fig. 5 while the X-axis means the characters and the Y-axis represents d of them. From left to right on the X-axis, d is reducing, it means the dissimilarity of dialect pronunciation to Mandarin are reducing. In Fig. 5, figure (a) shows the dissimilarity order of finals between Min speaker 13 and the female Mandarin speaker, figure (b) shows the dissimilarity to the male Mandarin speaker. Figure (c) shows the dissimilarity order between Min speaker 14 and the female Mandarin speaker, while figure (d) shows the dissimilarity to the male Mandarin speaker. Then by these figures, it can be found that the dissimilarity orders between the two Min speakers and the two Mandarin speakers are very similar. Take the six red characters /辣, 林, 耳, 桂, 資, 知/ as examples, they are all located at the left edge of all the results. It means that the Min and Mandarin pronunciation of these characters are extremely different and the results also show high invariance



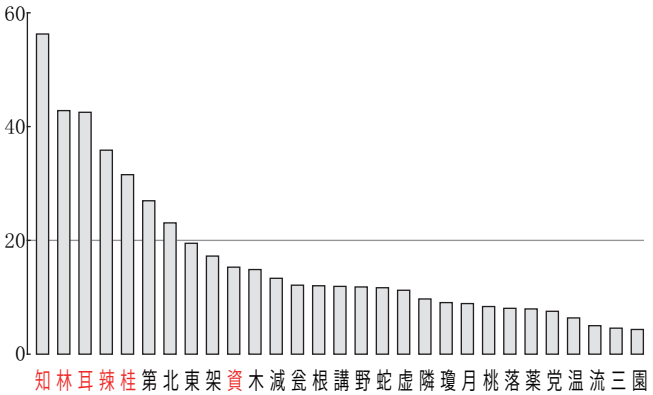
(a): Between Speaker 13 and the female Mandarin speaker



(b): Between Speaker 13 and the male Mandarin speaker



(c): Between Speaker 14 and the female Mandarin speaker



(d): Between Speaker 14 and the male Mandarin speaker

Fig. 5 Vowel similarity orders to standard Mandarin

to extra-linguistic features, the genders of the speakers.

6. Conclusions

In this paper, we proposed to apply a structural representation of Chinese dialects to pronunciation assessment. Using the utterances of different dialect speakers, the comparable dialect pronunciation is built for every speaker by extracting the purely linguistic features and canceling the variations caused by extra-linguistic features. Then with the utterances of 16 dialect speakers, some experiments are carried out to analyze their pronunciation by building the pronunciation structures and calculating the distances among them. After that, experiments of comparing the dialect pronunciation of individual utterances to standard Mandarin are done and the dissimilarity orders in their pronunciations are obtained. All the results show that our approach works well at assessing the pronunciation of dialect speakers linguistically and is highly independent of extra-linguistic variations caused by speaker variability. Therefore, this approach can be further applied to the pronunciation assessment of any speakers, no matter what kind of Chinese they are speaking.

References

[1] N.Minematsu, "Mathematical evidence of the acoustic universal structure in speech," ICASSP, pp. 889-892, 2005.

[2] N. Minematsu et al., "Theorem of the invariant structure and its derivation of speech gestalt," Int. Workshop on Speech Recognition and Intrinsic Variations, pp. 47-52, 2006."

[3] S. Asakawa et al., "Multi-stream parameterization for structural speech recognition," ICASSP, pp. 4097-4100, 2008.

[4] Y. Qiao et al., "f-divergence is a generalized invariant measure between distributions," INTERSPEECH, pp. 1349-1352, 2008.

[5] N. Minematsu et al., "Structural representation of the pronunciation and its use for CALL," Workshop on Spoken Language Technology, pp.126-129, 2006.

[6] D. Saito et al., "Structure to speech – speech generation based on infantlike vocal imitation –," INTERSPEECH, pp. 1837-1840, 2008.

[7] X. Ma et al., "Dialect-based speaker classification of Chinese using structural representation of pronunciation," SPECOM, pp. xxx-yyy, 2008.

[8] Yuan Jiahua et al., "HanYu FangYan GaiYao," Language & Culture Press, 2000.

[9] Hou Jingyi et al., "XianDai HanYu FangYan GaiLun," ShangHai Education Publishing House, 2002.

[10] M. Pitz et al., "Vocal tract normalization equals linear transformation in cepstral space," IEEE Trans. Speech and Audio Processing, vol. 13, no. 5, pp. 930-944, 2005.

[11] Z. Li, HanZi GuJin YinBiao, ZhongHua Book Company,1999.

[12] Richard VanNess Simmons et al, Handbook for Lexicon Based Dialect Fieldwork, Zhonghua Book Company, 2006.

[13] Institute of Linguistics of Chinese Academy of Social Sciences,Hanyu DiaoCha ZiBiao, The Commercial Press, 2007.

[14] Chinese Academy of Social Sciences, Language Atlas of China, Hong Kong: Longman Group, 1988