Structural Analysis of Dialects, Sub-dialects and Sub-sub-dialects of Chinese

Xuebin MA[1], *Akira NEMOTO*[2], *Nobuaki MINEMATSU*[1], *Yu QIAO*[1], *Keikichi HIROSE*[1]

[1]The University of Tokyo, Tokyo, Japan, [2]Nankai University, TianJin, China

{xuebin,mine,qiao,hirose}@gavo.t.u-tokyo.ac.jp, akiranmt@hotmail.com

Abstract

In China, there are hundred kinds of dialects. By traditional dialectology, they are classified into seven big dialect regions and most of them also have many sub-dialects and sub-sub-dialects. As they are different in various linguistic aspects, people from different dialect regions often cannot communicate orally. But for the sub-dialects of one dialect region, although they are sometimes still mutually unintelligible, more common features are shared. In this paper, a dialect pronunciation structure, which has been used successfully in dialect-based speaker classification in our previous work [1], is examined for the task of speaker classification and distance measurement among cities based on sub-dialects of Mandarin. Using the finals of the dialectal utterances of a specific list of written characters, a dialect pronunciation structure is built for every speaker in a data set and these speakers are classified based on the distances among their structures. Then, the results of classifying 16 Mandarin speakers based on their sub-dialects show that they are linguistically classified with little influence of their age and gender. Finally, distances among sub-sub-dialects are similarly calculated and evaluated. All the results show high validity and accordance to linguistic studies.

Index Terms: Sub-dialects of Mandarin, pronunciation structure, speaker classification

1. Introduction

In China, the current situation of its dialects is very complicated. As even people from two adjacent cites have difficulty in oral communication, standard Mandarin has been popularized all over the country as official language. Then for some socio-cultural reasons, the study of individual Chinese dialects and their relationships has become necessary and popular. These years, many results especially about the interrelationships of the dialects are published. In linguistic literatures [2] and [3], the affinity and mutual intelligibility among Chinese dialects are studied quantitatively based on their phonological structures. In [4] and [5], their distances are quantified based on individual kinds of phonological units like finals. But if one wants to classify speakers automatically based on their dialects and sub-dialects, and that using their utterances only, purely dialectal features should be extracted from the utterances.

In modern speech 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, and so on. So in order to capture the purely linguistic information from speech, the extra-linguistic aspects of speech should be removed, or if it is difficult, one has to ask human linguists to listen to and check the utterances for dialect-based speaker classification. In our previous study, in order to capture the purely linguistic information, structural representation of Chinese dialects was proposed and applied in classification of speakers based on their dialects and then, satisfactory results were obtained [1]. Meanwhile, as this structure is calculated by extracting speaker-invariant speech contrasts or dynamics and it shows high speaker independence, it was also applied in speaker-independent ASR [6] [7], speech synthesis [8] and Computer Aided Language Learning [9].

In this paper, the structural representation is used to calculate the interrelationships among sub-dialects of Mandarin by the acoustic features of their finals. In Section 2, the current situation of Chinese dialects and some phonological knowledge are introduced. After that, a pronunciation structure covering the dialectal changes of Chinese finals is proposed in Section 3. In Section 4, after the introduction of experimental data of sub-dialects of Mandarin, some experiments are described and the results are discussed with related linguistic knowledge. At last, this paper is concluded in Section 5.

2. Background of Chinese dialects

Generally, Chinese dialects are mainly grouped into 7 big dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue, Min) by traditional dialectology and each of them is divided into different sub-dialects and sub-sub-dialects [10]. For example, there are 8 sub-dialects in GuanHua dialect region and they are further classified into 42 sub-sub-dialects. As all these dialects are developed from the same root, many common features have been inherited. They are sharing the similar written characters, similar phonetic features and the same phonological structures. For example, every written character is pronounced as a monosyllable which is combined by an initial, a final and a tone, while the initial is always a constant and the final is mainly consisted of a vowel together with an optional coda. However, due to many historical or geographical reasons, there are still many differences among these dialects grammatically, lexically, phonologically and phonetically nowadays. Take the finals as example, there are 38 finals in Mandarin but 53 finals in Cantonese and 32 finals in Shanghainese. Therefore, people from different dialect regions cannot communicate orally sometimes. Further, for the sub-dialects of the same dialect region, although their phonological features are generally the same, there are still many differences to some degrees and people also have some difficulty in oral communication sometimes.

Because of the mutual unintelligibility among these dialects, Standard Mandarin has been popularized all over the country by the Chinese government as standard spoken language and almost every dialect speaker began to learn Mandarin. But affected by their native dialects, many of them speak Mandarin with accents to different degrees and some dialectal standard spoken languages have been generated in some big dialect regions, which are actually mixtures of standard Mandarin and native dialects. Meanwhile, because of the movement of

Table 1: Examples of selected characters

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

people and some other reasons, even for speakers from the same dialect region, their dialectal features are different sometimes.

3. Structural representation of dialects

3.1. Speaker-invariant dialect structures

When speech is represented by spectrum, its extra-linguistic information can be approximately modeled by two kinds of distortions: convolutional and linear transformational distortions. Microphone differences are the typical reason of convolutional distortions and vocal tract length differences are the typical reason of linear transformational distortions [11]. If a speech event is represented by cepstrum vector c, the convolutional distortion changes c into $c' = c + b$, meanwhile, the linear transformational distortion changes c into $c' = Ac$. So the total spectral distortions caused by extra-linguistic features can be modeled by $c' = Ac + b$, which is an affine transformation. So if we have an acoustic feature which is invariant to affine transformations, we can say that is invariant to extra-linguistic features.

Here, every speech event, e.g. a vowel is captured as a distribution ($p_i(c)$) and event-to-event distances are calculated as Bhattacharyya Distance (BD) because it is affine-invariant.

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

Then a distance matrix can be obtained by calculating BDs between any pair of all the events considered. Since a distance matrix can fix uniquely its geometrical shape composed of all the events, we call this distance matrix a structure. If vowel structures are extracted from the utterances of the same vowel set spoken by two speakers belonging to different dialects, the structures are highly expected to show only a dialectal difference not an extra-linguistic difference between the two speakers.

3.2. Comparable structures among dialects

In order to calculate the interrelationships among Chinese dialects, dialectal utterances of the same set of linguistic units are necessary, which must cover the differences among Chinese dialects sufficiently. In this paper, since we want to classify speakers based on the finals of their sub-dialects, sufficient finals of different dialects should be focused on. However, there is a problem that the final inventory changes according to dialects and they cannot be compared directly. However, as all the dialects have inherited the same written characters and every character has its final, we can compare the dialectal utterances using the same list of written characters and can measure the distances among these dialects.

Nowadays, the phonological features of modern Chinese dialects are always studied together with the historical phonology. 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 be compared. For example, the historical pronunciations and modern dialectal pronunciations of the commonly used written characters are all listed in [12]. Then based on these studies, some specific lists of written characters are always adopted by linguists to check the features of corresponding initials, finals and tones in different dialects [13] [14]. In [14], which is written by linguists in the Institute of Linguistics of Chinese Academy of Social Sciences, three different lists of written characters are listed 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 for every speaker can be built and the speakers of sub-dialects of Mandarin can be classified by calculating the distances among structures of different dialects. In our study, the list of written characters in [14], which is used for checking the dialectal finals, is adopted to build the comparable dialectal structures of individual speakers. In Table 1, some examples of these written characters and their corresponding syllables and finals are listed.

Table 2: Detailed information of the speakers

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	RongCheng	F

4. Experiments with dialect structures

4.1. Preparation of the experimental data

Using the selected written characters in Table 1, the recording was carried out in Nankai University in China. The subjects are 16 speakers from 8 cities belonging to 4 sub-dialects regions of Mandarin. 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. They are mainly undergraduate students of Nankai University and have no background of other languages before entering the university in Tianjin. For the following experiments, every speaker is given an ID which is listed in Table 2, together with the information about their hometown, their sub-dialect region and gender.

All the recordings were carried out in a quiet room with a supervisor, so the data are all expected to be clean. Before the recording, the Mandarin sub-dialectal pronunciation 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 sub-dialects of Mandarin four times. Then the data was labeled phonetically and manually by linguistic students. After checking the spectrum and raw file,

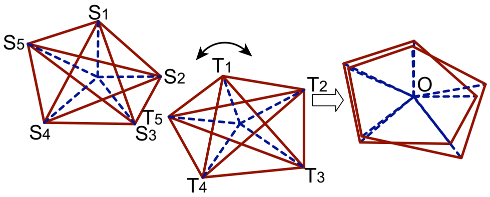

Figure 1: Distance calculation after shift and rotation

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

every syllable was labeled into two parts, initial and final, with transcriptions mainly developed from Chinese Pinyin.

4.2. Speaker classification based on structures

The final part of every syllable is modeled as a single Gaussian distribution under the acoustic conditions shown in Table 3. Then, for every speaker, the BDs of every pair of finals are calculated to form his/her dialect pronunciation structure, which is expected to show all the dialectal features of the final utterances of him/her. Then these speakers can be classified by calculating the distances among their pronunciation structures.

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. 1. With the best overlap, the minimum sum of the distances between the corresponding two points of the two structures can be obtained. In [15], it was experimentally proved that the minimum sum can be approximately calculated as Euclidean distance between two distance matrices by the following formula:

$$D(S,T) = \sqrt{\frac{1}{M}\sum\nolimits_{i<j}(S_{ij}-T_{ij})^2}, \qquad (2)$$

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

4.3. Speaker classification based on dialects

In our previous work [1], speaker classification based on dialects, not sub-dialects, were investigated especially in terms of robustness to speaker variability. After the data of 18 speakers of 4 dialects were recorded, simulated data of tall and short speakers were also obtained by applying a frequency warping technique [11] to the original data. Then the simulated speakers and the original speakers, the number of whom is 54 in total, were classified by our method and the conventional method. Fig. 2 and Fig. 3 are the results. Fig. 2 was obtained by using D_1, but Fig. 3 was obtained by directly and acoustically comparing the spectrums between speakers, which is often done in DTW. In these figures, every speaker is presented by an ID, while the ID with a line on the top represents the simulated tall speaker and the ID with a line on the bottom represents the simulated short speaker. Besides, the color means their dialect regions and IDs in italic type means they are female. In Fig. 2, the speakers from the same dialect region are all clustered together and shows high independence of the gender and other extra-linguistic factors, because the simulated tall and short speakers

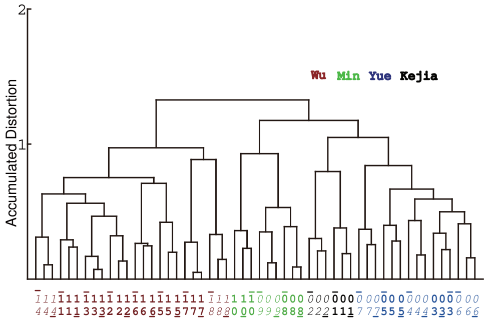

Figure 2: Speaker classification using our approach

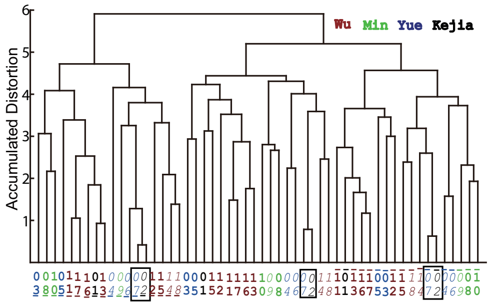

Figure 3: Classification using the conventional approach

are all clustered together with the original ones. In Fig. 3, although using the same data, the speakers are classified into three big sub-trees corresponding to their vocal tract length with no relation to their dialects. Besides, in Fig. 3, 02 and 07 are very closely clustered. They are the same speaker who can speak two dialects of Kejia and Yue. In Fig. 2, they are different and in Fig. 3, they are almost the same.

4.4. Speaker classification based on sub-dialects

In this paper, by using the proposed method, we challenge speaker classification based on the sub-dialects of Mandarin. The result with the new data of 16 speakers is shown in Fig. 4. The ID of every node is the same as that in Table 2 and the colors mean different sub-dialect regions. In this figure, the speakers are mainly classified by their sub-dialects and the speakers from the same city are all classified together. The 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 also well clustered to two sub-trees respectively. For the speakers from JiLu sub-dialect region, although speakers 05-06 from YuZhou and speakers 07-08 from ShangQiu are still grouped together 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 [16]. According to [16] and [10], 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 our experiments, only the finals are adopted and their pronunciation of the finals with nasal coda

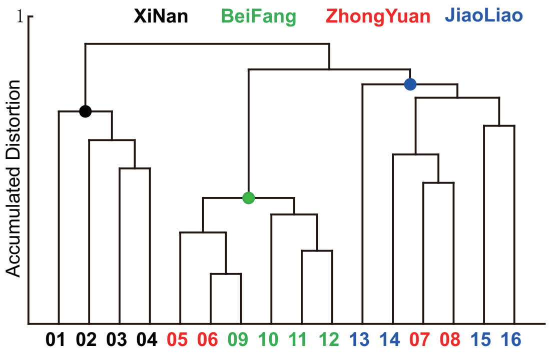

Figure 4: Classification of the speakers based sub-dialects

are generally the same in these three sub-dialect regions. Therefore, the speakers 05-16 are all clustered in a big sub-tree and speakers 05-08 are clustered to the neighboring trees. In conclusion, this result proves that dialect pronunciation structure can also work well on extracting the purely linguistic information of sub-dialects of Mandarin.

4.5. Distances among hometowns based on dialects

In the above experiments, the dialect pronunciation structure is proved to work well at capturing speaker-invariant features of dialects and sub-dialects. Then, we will discuss applying the structure to calculating the distances among hometowns based on the sub-dialects. Strictly speaking, although two cities belong to the same sub-dialect region, the pronunciation somewhat differs between them. Here, the 8 hometowns of the 4 sub-dialects are considered to stand for 8 sub-sub-dialects. By building the pronunciation structure through averaging the structures of the speakers belonging to each sub-sub-dialect (hometown), the inter-town distances are calculated. The result is shown in Fig. 5, 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 [17], 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 acoustic distances among dialects can be obtained and it could be a good and objective proof of the study of linguists.

5. Conclusions

In this paper, a novel approach of using dialect pronunciation structure, which can extract the speaker-invariant dialect features from speech, is proposed to sub-dialect based speaker classification and measurement of acoustic distance among sub-sub-dialects. After a common list of written characters is selected, a dialect pronunciation structure is built for every speaker using their dialectal utterances of these characters. Then these speakers are classified based on the distances among these structures and satisfactory classification results are obtained. At last, using the dialect structures of these speakers, the dialect structures for these hometowns are obtained and their acoustic distances are calculated and discussed. Besides, we have just finished collecting dialectal utterances from 66 speakers and the final results using them will be presented in the conference.

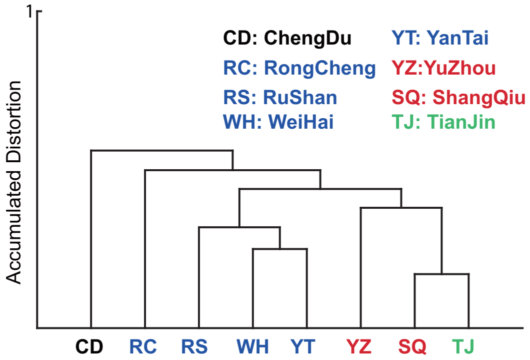

Figure 5: Classification of the hometowns

6. Acknowledgements

The authors of this paper would like to thank the teachers and students of College of Chinese Langauge & Culture in Nankai for their help of data post-process and verifications.

7. References

[1] X. MA, et al,"Dialect-based Speaker Classification of Chinese Using Structural Representation of Pronunciation", SPECOM, 2009.

[2] C. Cheng, "Syllable-based dialect classification and mutual intelligibility", Symposium Series of the Institute of History and Philology, Academia Sinica Number 2, Chinese Languages and Linguistics I Chinese Dialects 145-177, 1992.

[3] C. heng, "Measuring relationship among dialects: DOC and related resources", International Journal of Computational Linguistics & Chinese Language Processing 2.1:41-72, 1997.

[4] H. Chen, "Measurement of the similarity among dialect finals systems", Zhongguo Yuwen, 275, pp.139-145, 2000.

[5] H. Chen, "Calculation of phonological similarity between dialects", Language Sciences, 5.1, pp.23-31, 2006.

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

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

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

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

[10] J. Yuan et al, HanYu FangYan GaiYao, Language & Culture Press, 1998.

[11] 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.

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

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

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

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

[16] J. Hou et al, XianDai HanYu FangYan GaiLun, ShangHai Educational Press, 2002.

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