Dialect-Based Speaker Classification Using Speaker-Invariant Dialect Features

Xuebin Ma
Graduate School of Frontier Sciences,
The University of Tokyo,
Tokyo, Japan
Email: xuebin@gavo.t.u-tokyo.ac.jp

Ruiyuan Xu
Institute of linguistics,
Chinese Academy of Social Sciences,
Beijing, China
Email: xury@cass.org.cn

Nobuaki Minematsu
Graduate School of Information
Science and Technology,
The University of Tokyo,
Tokyo, Japan
Email: mine@gavo.t.u-tokyo.ac.jp

Yu Qiao
ShenZhen Institute of Advanced Technology,
Chinese Academy of Science,
ShenZhen, China
Email: yu.qiao@sub.siat.ac.cn

Keikichi Hirose
Graduate School of Information
Science and Technology,
The University of Tokyo,
Tokyo, Japan
Email: hirose@gavo.t.u-tokyo.ac.jp

Aijun Li
Institute of linguistics,
Chinese Academy of Social Sciences,
Beijing, China
Email: liaj@cass.org.cn

***Abstract*—In our previous works, a structural pronunciation representation was proposed to extract the linguistic features from dialect pronunciation and classify speakers based on their dialects. In this paper, in order to prove that the structural method can extract the purely speaker-invariant dialectal features, several new experiments are carried out. First, using the data of 19 speakers from different dialect and sub-dialect regions, a dialect-based speaker classification experiment is carried out and satisfactory result is achieved. Then, one Chinese dialectologist transcribes all the data and reads the linguistic content of each original utterance in her voice through looking at the transcript and listening to the original utterance. So a new data set with minimum speaker differences (fixed speaker identity) is created. Using the new data, similar classification experiment is carried out and the result is very similar to the result of last experiment. It means that our method can extract the purely speaker-invariant dialectal features and classify speakers based on their dialects very well. After that, for the original and mimicked data sets, data sets with maximum speaker differences are simulated using high-quality voice morphing techniques. Using the original dialect data and the simulated versions together, classification experiments are carried out based two criteria, spectral comparison and structural comparison. By comparing these results, we can find that unlike the method of spectral comparison, the structural method can purely classify speakers based on their dialects, which shows the proposed dialect structures are speaker-independent and linguistic enough features.**

I. Introduction

Generally speaking, dialect means a variety of a language that is used by a particular group of that language's speakers. Among dialects, there are always some phonetic, grammatical, and lexical differences to different degrees. 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, and so on. Therefore, in order to process different dialects in conventional spectrum-based dialect processing frameworks, dialect-dependent but speaker-independent models were always trained by collecting utterances from many different speakers of one dialect. However, this approach may not work well especially in Chinese dialect processing.

In China, there are hundreds kinds of dialects. Traditionally, they are classified into 7 major dialect regions [1] and every dialect region has many sub-dialects and sub-sub-dialects [2]. Therefore, in order to process all the sub-dialects of one dialect region by training different models for different sub-dialects, dozens of models must be built sometimes. Further, because of the popularization of Mandarin and population movement across different dialect regions, the dialects of many speakers are also changing. So two speakers from the same dialect region may speak different sub-dialects and it is a very challenging work to build dozens of sub-dialect models for one dialect region by collecting the data of many speakers from the same sub-dialect region.

In our previous study, a structural pronunciation representation was proposed to extract speaker-invariant speech contrasts or dynamics [3], [4] and applied to speaker-independent Automatic Speech Recognition (ASR) [6], speech synthesis [7] and Computer Aided Language Learning (CALL) [8]. Then, this approach was further applied to Chinese dialect analysis [9] and dialect-based speaker classification [10] with satisfactory results were achieved.

In this paper, the dialect pronunciation structure is applied to extracting the purely linguistic features from Chinese dialect to classify speakers based on their dialects and the speaker-invariance of this approach is examined. In Section 2, the current situation and fundamentals of Chinese dialects are introduced. Then the method for building comparable dialect pronunciation structures and calculating the distance between them is described in Section 3. In Section 4, dialect-based speaker classification experiment is carried out using the dialect data of 19 speakers. In Section 5, this proposal is verified by classification experiment using data with minimum speaker differences. In Section 6, this proposal is further verified by classification experiment using data with maximum speaker differences and the result is compared with the classification based on spectral comparison. At last, this paper is concluded in Section 7.

TABLE I
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/

TABLE II
DETAILED INFORMATION OF 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

II. FUNDAMENTALS OF CHINESE DIALECTS

In China, there are hundreds kinds of dialects and they are traditionally classified into 7 major dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue and Min) [1]. Moreover, most of the major dialects also have many different sub-dialects and sub-sub-dialects. For example, there are 8 sub-dialects and 42 sub-sub-dialects in GuanHua dialect region. All the dialects are developed from the same root 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 example, every written character is pronounced as a monosyllable which is combined by an initial, a final and a tone. However, due to many historical or geographic reasons, there are still many differences among these dialects grammatically, lexically, phonologically and phonetically. Take the finals as example, there are 38 finals in Mandarin but 53 finals in Cantonese and 32 finals in Shanghainese.

Since 1956, standard Mandarin has been popularized all over the country as 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 regional accents to different degrees and their native dialects also start changing affected by Mandarin. In addition, because many people of different dialect regions are moving all over the country, the dialect of individual speakers is also changing affected by different language backgrounds.

In brief, the current situation of Chinese dialects is becoming more and more complicated. Strictly speaking, every speaker has his/her own dialect, and the pronunciations of two speakers of the same dialect often show different sub-dialect features because they may belong to different sub-sub-dialects.

III. PRONUNCIATION STRUCTURE OF DIALECTS

A. Mathematical model of extra-linguistic information

When speech is represented acoustically by spectrum, the inevitable extra-linguistic features can be approximately modeled as two kinds of distortions according to their 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 for linear transformational distortions [11]. If a speech event is represented by a 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 a 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 the affine transformation.

B. Speaker-invariant structure in dialects

Here, every speech event, such as the pronunciation of one syllable, is captured as a distribution and event-to-event distances are calculated as Bhattacharyya Distance (BD),

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

where $p_1(c)$, $p_2(c)$ mean the distributions of two speech events. With multiple events, we can obtain a distance matrix by calculating BDs between any pair of them. Since BD is invariant with respect to affine transformations, the obtained matrix is invariant to extra-linguistic factors. As a distance matrix can determine uniquely a geometric shape, we refer to the matrix as a pronunciation structure. Therefore, with the utterances of dialect speakers, we can build dialect pronunciation structures which are invariant to extra-linguistic factors.

C. Comparable dialect pronunciation structures

In order to analyze the pronunciation of speakers from different dialects using the structural representation, comparable dialect structures have to be built using their dialect utterances of the same set of some linguistic units. Considering that although there are many grammatical and lexical differences and the inventory of the phonological units changes from dialect to dialect, all the Chinese dialects share the same

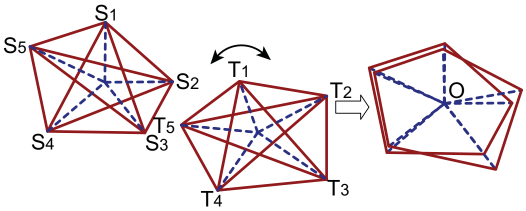

Fig. 1. Distance calculation after shift and rotation

written characters and every character is pronounced as a mono-syllable, the utterances of syllable units (characters) become the best choice to build the comparable structures.

Recently, some Chinese linguists are focusing on the relationships among the dialects by their phonological features. According to their studies, some specific lists of characters are proposed to check the phonological differences among dialects. For example, in [12], three different lists of characters are shown for checking the dialect features of tones, initials and finals, separately. Therefore, using these lists, different comparable dialect pronunciation structures can be built to check different features. In order to classify speakers of different dialects, the characters covering all the dialect differences can be adopted. With the dialect utterances of these characters, a speaker-invariant but dialect-sensitive pronunciation structure can be built for every speaker and speakers can be classified based on their dialects by calculating the distances among the structures.

In our study, the list of written characters in [12], which is used for checking the finals, is adopted to build the comparable dialectal structures of individual speakers. Some examples of these written characters and their corresponding syllables and finals of standard Mandarin are listed in Table I.

D. Distances between pronunciation structures

After the dialect pronunciation structures were built for the speakers using the BD among their utterances, the distances among their dialects can be calculated as 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 like in Fig. 1. In [3], it was experimentally proved that this distance can be approximately calculated as Euclidean distance between two structures. Following is the detailed computing formula:

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

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

IV. Classification experiment using original data

A. Experimental data of dialects

We found that publicly available Chinese dialect corpora cover only two or three dialects and cannot be used for our purpose. Then we carried out some recordings for our experiments. The reading materials is the list of written characters in [12]. The recordings were carried out at a university in China and the recording subjects were all university students. Totally, 19 speakers joined our recordings and they belonged to 10 different sub-dialect regions from 5 general dialect regions. Every speaker was given an ID and more information such as the sub-dialect regions and genders of them can be found in Table II. All the recordings were carried out in quiet rooms with a supervisor. Every speaker was asked to read the selected characters in their native dialects three times. Then after all the data were labeled phonetically by students of linguistics, the final of every syllable was modeled as a single Gaussian distribution under the acoustic conditions shown in Table III. After that, for every speaker, the BDs between his/her utterances are calculated and the dialect pronunciation structure is built.

TABLE III
Acoustic analysis condition

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

B. Experiment using the original dialect data

Using D_1, the distance between the dialects of two speakers is calculated as the distance between the pronunciation structures of them. Then these speakers are classified and the result is shown by Fig. 2, where the structure of every speaker is represented by the speaker ID in Table II and different colors show different dialect regions. In this figure, the result is shown by a bottom-up clustering method, Ward's clustering method.

In this figure, 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 II, we can find that the speakers from the same sub-dialect are all classified near to each other in the result. But looking at speakers G3 and H2, we have to admit that all the speakers are not completely clustered into different sub-sub-trees by their dialects. Then, a question will still be asked that whether there is any problem with our approach or just these speakers should be classified in that way according to their acoustic features.

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-dialects of Hakka are located at the middle of Gan dialect region and the Xiang dialect region are also very close to Gan dialect region geographically [2]. And about speaker G3 and H2, before the experiment, their data were checked by a dialectologist. It is found that the dialects of G3 and H2 are most different to other Gan speakers and Hakka

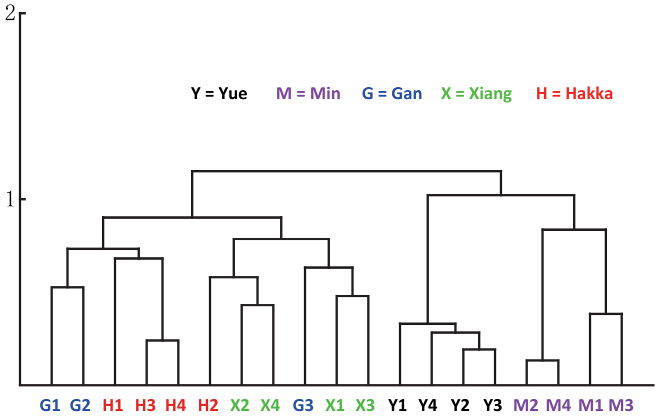

Fig. 2. Classification result using the original dialect data

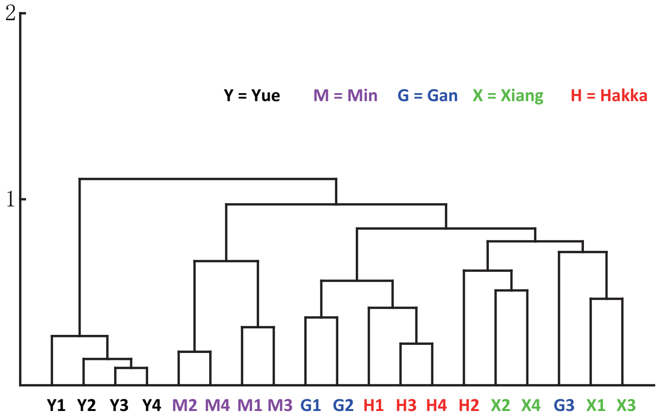

Fig. 3. Classification result using the new mimicked data

speakers, respectively, and their three pronunciations of some characters are not very steady. Meanwhile, it is also considered that the linguistic distances of these dialects are different to their acoustic distances because traditional linguists classify Chinese dialects not only according to their acoustic features. Thus, the acoustic distances of these dialects and the linguistic distances of them cannot be compared directly. So we design a new experiment to prove our approach can extract the purely speaker-invariant linguistic or dialect features.

V. Verification Experiment with Data of Minimum Speaker Differences

A. Linguistically mimicked data

For the new experiment, we tried to build a new corpus that the features of speaker differences are removed manually. In fact, if there is a Chinese dialectologist who can speak all these dialects, the dialect utterances recorded above can be repeated linguistically by him/her, which gives us the dialect utterances only with a fixed speaker identity.

In fact, nobody can speak all the Chinese dialects. But 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. At last, the second author finished this challenging work. Then, the new version of data was checked at least twice by different linguists. By listening to the original utterance and the corresponding new one, they were ensured to be the same linguistically.

Using the mimicked data of multi-dialects but fixed speaker identity, a verification classification experiment can be carried out. If the classification result is similar to the result in Fig. 2, it will mean that our approach can extract the purely linguistic features by canceling the features of speaker differences.

B. Classification experiment using mimicked data

Using the mimicked data, the classification experiment is carried out and the result is shown in Fig. 3, while the IDs and colors are the same as those in Table II. By comparing this result with Fig. 2, it was found that they are very similar to each other. In both of these results, all the speakers are classified into four large sub-trees and each has the same speakers: speakers from Yue and Min are classified into their individual sub-trees; speakers from Xiang are also classified into a large sub-tree and speaker H2, G3 are also classified into this sub-tree as well; the left Gan speakers and Hakka speakers were clustered into a large sub-tree, which itself has two sub-sub-trees corresponding to Gan and Hakka separately. By focusing on the speakers (Gan, Hakka and Xiang), we can find their positions are exactly the same in the two results. So it means that our approach can be utilized to extract the speaker-invariant purely linguistic features from the dialect pronunciation of every speaker.

VI. Verification Experiment with Data of Maximum Speaker Differences

A. Simulated data of speakers with long and short vocal tracts

It is known that the vocal tract length of speaker is an important extra-linguistic feature and rotates a utterance trajectory in cepstrum space [13]. Generally speaking, 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, 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 $(=\{a_{ij}\})$ [11].

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0,j-i)}^{j} \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!} (-1)^{(m+i-j)} \alpha^{(2m+i-j)} \tag{3}$$

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. When $\alpha > 0$, formants are transformed to be higher and the vocal tract length shorter. Considering the height of the world tallest adult and shortest adult, $\alpha = 0.2$

and $\alpha = -0.2$ can be used to create the data of the tallest and shortest speaker. Using matrix A, the original utterances and mimicked data were converted into a shorter version with $\alpha = 0.2$ and a taller version with $\alpha = -0.2$ using a high-quality analysis-resynthesis system, STRAIGHT [14]. We can regard these data as the data with maximum speaker differences.

B. Spectral classification using the simulated data

In the conventional acoustic matching framework 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:

$$D_2(S,T) = \sqrt{\frac{1}{M}\sum_{i} BD(F_i^S, F_i^T)}. \tag{4}$$

F_i^S is syllable utterance i of speaker S and F_i^T is utterance i of speaker T. M means the number of the utterances.

Using the original dialect data and the simulated versions, a classification experiment was carried out by calculating the spectral distances between them using D_2. The result is shown by Fig. 4. The speaker IDs and the colors have the same meanings as in Table II, while an ID with a top bar means a simulated taller speaker and an ID with a bottom bar means a simulated shorter speaker. In this figure, we can find that speakers are classified into three big sub-trees according to their body heights. And in each sub-tree, the classification is affected by the speaker features greatly and many speakers are not classified by their dialects.

C. Structural classification using the simulated data

Using the mimicked data and the warped utterances, their dialect pronunciation structures are built and speakers are classified based on the distances between them and the result is shown in Fig. 5. In this result, the speaker IDs and the colors are the same as those in Table II, while an ID with a top bar means a simulated taller speaker and one with a bottom bar means a simulated shorter speaker.

In Fig. 5, it is found that all the speakers are classified by their dialects and the simulated tall and short speakers are classified near to their corresponding original speaker 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. 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. 3 which is obtained only using the original dialect data.

In brief, we can find the structural method still works very well even using the dialect data with minimum speaker differences (mimicked data) and maximum speaker differences (simulated data) together. Unlike the classification using conventional spectral comparison, these speakers are classified by their dialects and the result is not affected by speaker features at all. It is further proved that our structural method can extract the purely dialect features from speech.

VII. Conclusions

In this paper, several experiments are carried out and show that the structural representation of Chinese dialect pronunciations can extract the speaker-invariant linguistic features and classify speakers based on their dialects. At the beginning, a dialect-based speaker classification experiment is carried out using the utterances of 19 dialect speakers. Then the original utterances spoken by different speakers are read linguistically by one experienced dialectologist in her own voice and a new corpus with minimum speaker differences are built. Using the mimicked data, classification experiment is carried out and the result is very similar to the result obtained using original dialect data. After that, data sets with maximum speaker differences are built using high-quality voice morphing techniques and several verification experiments are carried out using our structural comparison and the conventional spectral comparison. By these results, our proposal is shown again that it can extract the purely linguistic features and the classification result are not affected by the speaker features.

Acknowledgment

The first author of this paper would like to thank M.S. HongLi Liang and several other students of the Phoetic Lab of CASS for their help of labeling and checking the experiment data.

References

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

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

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

[4] 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.

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

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

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

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

[9] X. Ma et al., "Structural analysis of dialects, sub-dialects, and sub-sub-dialects of Chinese," Proc. INTERSPEECH, pp.2219-2222, 2009.

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

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

[13] Saito, D., Minematsu, N. and Hirose, K., "Decomposition of rotational distortion caused by VTL difference using eigenvalues of its transofmation matrix", INTERSPEECH, pp. 1361-1364, 2008.

[14] 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.

[This paper was published in Speech Prosody 2010]