Structural Analysis of Chinese Dialect Speakers and Their Automatic Classification

XueBin Ma[1], Nobuaki Minematsu[2], Akira Nemoto[3], Max Takazawa[4], Yu Qiao[2], Keikichi Hirose[2]

[1]Graduate School of Frontier Sciences, The University of Tokyo, Japan
[2]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[3]College of Chinese Language & Culture, Nankai University, Tianjin, China
[4]Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
{xuebin,mine,max,qiao,hirose}@gavo.t.u-tokyo.ac.jp, akiranmt@hotmail.com

Abstract: In China, there are many different kinds of dialects and sub-dialects. Because there are many grammatical, lexical, phonological, and phonetic differences among them in varying degrees, people from different dialect regions always have difficulties in oral communication. Since 1956, standard Mandarin has been popularized all over the country as official language and almost every dialect speaker began to learn Mandarin just as a second language. But affected by their native dialects, many of them speak Mandarin with regional accents. In modern speech processing technologies, speech is represented by spectrum which contains not only the dialectal linguistic information but also extra-linguistic information such as the gender and age of the speaker. In order to focus exclusively on the linguistic features of dialectal utterances, a speaker-invariant structural representation of speech, which was originally proposed by the second author inspired by infants' language acquisition [1, 2], is proposed to represent the pronunciation of Chinese dialect speakers. Since the purely dialectal information can be extracted by removing the extra-linguistic information from dialect speech, this pronunciation structure can be applied to estimate which dialect or sub-dialect region a speaker belongs to and to assess the pronunciation. In order to testify the validity of our approach, speaker classification based on the dialectal utterances of 38 Chinese finals are investigated especially in terms of robustness to speaker variability. The result is linguistically reasonable and highly independent of age and gender. After that, a sub-dialect corpus is developed with a list of characters as reading materials, which is originally used for linguists' investigation of dialect speakers' pronunciation. Then after the sub-dialect pronunciation structure is built for every speaker, their pronunciations are classified based on the distances among their structures. The result shows that the sub-dialect speakers can also be linguistically classified with little influence of their age and gender. In conclusion, this structural representation of Chinese dialects can extract the purely dialectal and sub-dialectal information from speech and works well on dialect-based and sub-dialect-based speaker classification.

Index Terms: Chinese dialects, extra-linguistic feature, pronunciation structure, Bhattacharyya distance, speaker classification

Segmental features of speech are usually represented acoustically by spectrum in modern technologies, which contains not only linguistic information but also extra-linguistic information corresponding to age, gender of speakers and so on. In other words, the same linguistic content is acoustically realized differently from a speaker to another. In the case of dialect pronunciation assessment, we should focus on the acoustic features of speech which is relevant to dialectal differences and irrelevant to extra-linguistic differences. 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 that the acoustic differences between a Mandarin utterance of an adult and its dialectal version of that adult. Therefore, in the case of automatic speech recognition (ASR) and computer-aided language learning (CALL), for each dialect, speaker-independent acoustic models are often built by collecting utterances from thousands of different speakers of this dialect but speaker adaptation or normalization techniques are still required. This approach, however, doesn't work well sometimes because, strictly speaking, speakers of the same dialect are often speakers of different sub-dialects.

As is known, infants acquire spoken language through imitating their parents' utterances but no child tries to produce their parents' voices. In fact, their

phonemic awareness is very immature and they cannot speak by imitating the individual speech sounds produced by their parents. Therefore, it is claimed by developmental psychology that they firstly acquire the holistic sound pattern of a (word) utterance and then, the segmental sound categories are learnt. In [3], the sound pattern is called as word Gestalt and it can be considered as the skeleton of a spoken word. And the word Gestalt must be speaker-invariant because children don't change their voice quality whoever talks to them. Inspired by this, a speaker invariant structural presentation was proposed to remove extra-linguistic and irrelevant acoustic features from utterances in our previous work [1, 2]. As this structure is calculated by extracting speaker-invariant speech contrasts or dynamics and shows high speaker independence, it can be viewed as speech Gestalt. Now, this speech structure was already applied in speaker-independent ASR, which was realized only with a small number of training speakers, where explicit speaker adaption or normalization was not needed [4, 5]. Further, the structure was also applied for helping Japanese learning English [6] and speech synthesis [7] with satisfactory results obtained.

In this paper, the pronunciation structure is applied to represent Chinese dialect and classify speakers based on their dialects. In Section 2, the current situation of Chinese dialects is introduced. In Section 3, the dialect sensitive but speaker-invariant speech structure is described. In Section 4, after the introduction of the experimental data of sub-dialects of Mandarin, some experiments are described and the results are discussed from a linguistic viewpoint. At last, this paper is concluded in Section 5.

1 The current situation of Chinese dialects

In China, there are many kinds of dialects and they are mainly grouped into 7 big dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue, Min) [8]. Further, most of them have some different sub-dialects and sub-sub-dialects too. For example, Guanhua (Mandarin) region can be grouped into 8 sub-dialects and many sub-sub-dialects [9]. Nevertheless, all these dialects and sub-dialects are developed from Old Chinese and Middle Chinese, and a lot of common features are inherited. Most of them share the same written scripts, very similar sound systems, the same phonological and structural features and so on. Take phonological features for example, every character is pronounced as mono-syllable with the same syllable structure which is combined by a tone, an initial and a final. The initial is always a consonant while the final is mainly consisted of a vowel. Among these dialects, however, there are still many differences grammatically, lexically, phonologically and phonetically. Even for the people from two adjacent cities, their dialects are sometimes different and they have difficulty in oral communication. Since 1956, standard Mandarin, the main branch of GuanHua dialect region, has been popularized all over the country as official language with the name of Putonghua. Then, almost every dialect speaker began to learn Mandarin just like learning a second language. However, many of them speak Mandarin with some regional accents affected by their native dialects. Generally, one can guess their native dialects easily according to their accented Mandarin if he/she has some knowledge of these dialects. On the other hand, as standard Mandarin is becoming more and more popular and many people of different dialect regions are moving all over the country, some dialects are losing some of their own unique features. Nevertheless, these dialects, especially some major dialects, are still widely used. And even outside their native dialect regions, people from the same dialect region always like to speak their own dialect to each other to show the special close relationship between them.

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 show somewhat different linguistic features because they may belong to different sub-dialects. To realize dialect-based speaker classification, it is necessary to use the dialectal features of speakers extracted through removing extra-linguistic features.

2 Estimation of dialect structures

2.1 Modeling the extra-linguistic speech variations

When speech is represented acoustically by spectrum, the inevitable extra-linguistic factors can be approximately modeled by two kinds of distortions according to their spectral behaviors: convolutional and linear transformational distortions. Convolutional distortions are caused by extra-linguistic factors such as

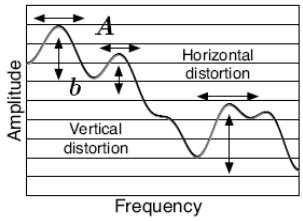

Fig. 1 Spectral distortions caused by Matrix A and vector b

different recording microphones, and vocal tract length differences are the typical reason of linear transformational distortions [10]. 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 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 distortions are schematized by Fig.1, where the horizontal and vertical distortions correspond to the distortions due to matrix *A* and vector *b*, respectively.

2.2 Speaker-invariant structures in dialects

As extra-linguistic variation in speech is modeled as affine transform, to obtain speech features invariant to extra-linguistic variation, we have to use affine-invariant features. In [9], Bhattacharyya Distance is shown to be invariant with affine transform.

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

Therefore, after every speech event is captured as a distribution and a distance matrix is obtained by calculating the BDs between any pair of speech events, this matrix becomes invariant to extra-linguistic variations. Here, we call this matrix a pronunciation structure of these speech events, because a distance matrix can represent uniquely its geometrical shape composed of all the speech events. Three examples of the invariant underlying structures are shown in Fig. 2. Any two of them can be converted to one another by multiplying matrix *A*. Although they look very different to each other, their BD-based distance matrices are identical.

Then the structural representation of a dialect speaker is 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 a single speaker who can speak different dialects, the difference will be large.

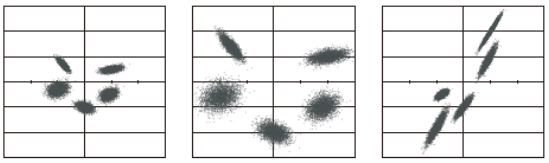

Fig. 2 Invariant underlying structure among three data sets

3 Comparison of dialect structures

3.1 Comparable structures among dialects

In order to evaluate the pronunciations of dialect speakers using the structural representation, comparable dialectal structures should be built with their dialectal utterances of the same set of linguistic units, which must cover the differences among Chinese dialects sufficiently. Then syllable or smaller phonological units become a good choice considering there are many grammatical and lexical differences among Chinese dialects. However, although all Chinese dialects are sharing the same phonological structures, the inventories of their phonological units are different and they cannot be compared directly. Nevertheless, since 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 structures to classify the pronunciation of speakers from different dialects. Therefore, if we can select a common list of characters which covers most of the phonological units in all the dialects, reasonable and comparable pronunciation structures for the dialects can be built and the pronunciation of different dialect speakers can be evaluated.

In these years, many Chinese linguists are studying Chinese dialects and their phonological features are always studied together with historical phonologies. 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

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/

characters are all listed in [11]. 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. Then the dialect pronunciation of every speaker can be assessed after all the speakers are classified by calculating the distances among their structures. Here, a list of written characters in [13], which is used for checking the dialectal finals, is adopted to build the comparable dialectal structures of dialect speakers individually. In Table 1, some examples of these written characters and their corresponding Mandarin syllables and finals are listed.

3.2 Measurement of distance between structures

Using the dialectal utterances of the selected characters, the pronunciation structure for every speaker can be built by the BDs of every pair of utterances, which is expected to show all the dialectal features of his/her pronunciation. Then by calculating the distances among their pronunciation structures, all the speakers can be classified based on their dialects. 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 the minimum sum can be approximately calculated as Euclidean distance between two distance matrices by the following formula:

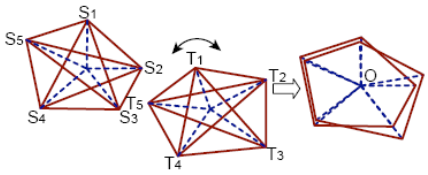

Fig. 3 Distance calculation after shift and rotation

Table 3: Acoustic analysis condition

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

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

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 utterances.

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 China and two kinds of subjects joined the recording. The first kind of subjects is 16 speakers from 8 cities belonging to 4 sub-dialect regions of Mandarin. They are mainly 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. The second kind of subjects is 2 adults at the age of 50 and 6 children at the age of 11 or 12. They were born and have been living in a small village which is located at the middle region of BeiFang and JiaoLiao sub-dialect regions. The 2 adults never learned Mandarin and the 6 children are learning Mandarin in the same class. 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 quiet rooms with a supervisor, so the data are all expected to be clean. Before the recording, the Mandarin sub-dialectal pronunciations of all the reading characters were checked by every speaker. Then the recording was carried out with a 48KHz linear PCM

recorder

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	ZhongYuan	YuZhou	F
06	ZhongYuan	YuZhou	F
07	ZhongYuan	ShangQiu	F
08	ZhongYuan	ShangQiu	F
09	JiaoLiao	YanTai	F
10	JiaoLiao	RuShan	M
11	JiaoLiao	WeiHai	F
12	JiaoLiao	RongCheng	F
13	BeiFang	TianJin	F
14	BeiFang	TianJin	M
15	BeiFang	TianJin	M
16	BeiFang	TianJin	F
17	Middle Area	LinQu	F
18	Middle Area	LinQu	M
19	Middle Area	LinQu	F
20	Middle Area	LinQu	F
21	Middle Area	LinQu	M
22	Middle Area	LinQu	M
23	Middle Area	LinQu	F
24	Middle Area	LinQu	M

of Sony PCM-D1. Every speaker was asked to read the selected characters in their native sub-dialects of Mandarin four times. All the data were labeled phonetically and manually by linguistic students. After checking the spectrum and raw file, 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.

4.2 Speaker classification based on dialects

In our previous work [14], using structural representation of dialects, speaker classification based on their dialects was 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 [10] to the

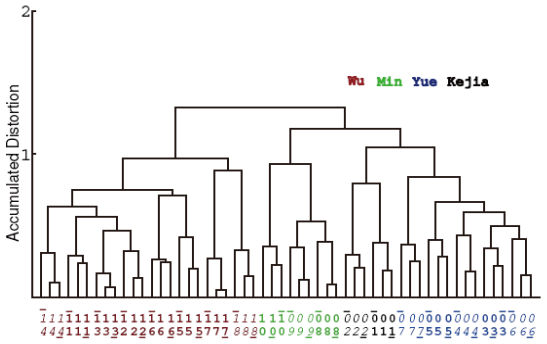

Fig. 4 Pronunciation classification using our approach

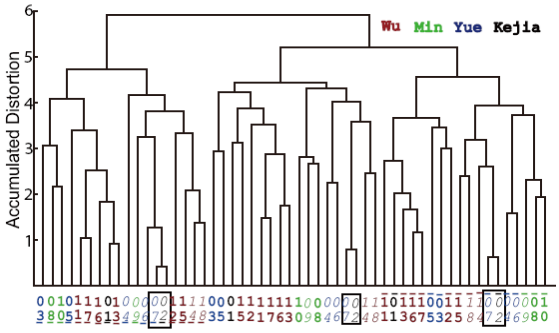

Fig. 5 Pronunciation classification by conventional approach

original data. Then the simulated speakers and the original speakers, the number of whom is 54 in total, were classified all together by our method and the conventional method. Fig. 4 and Fig. 5 are the results. Fig. 4 was obtained by using D1, and Fig. 5 was obtained by directly and acoustically comparing the spectrums between speakers. In these figures, every speaker is represented 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 colors mean their dialect regions and IDs in italic type mean they are female. In Fig. 4, the speakers from the same dialect region are all clustered together and the result shows high independence of the gender and other extra-linguistic factors, because the simulated tall and short speakers are all clustered together with the original ones. In Fig. 5, 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, 02 and 07 are the same speaker who can speak two dialects of Kejia and Yue. In Fig. 4, they are clustered into different dialect region, but in Fig. 5, they are clustered next to each other.

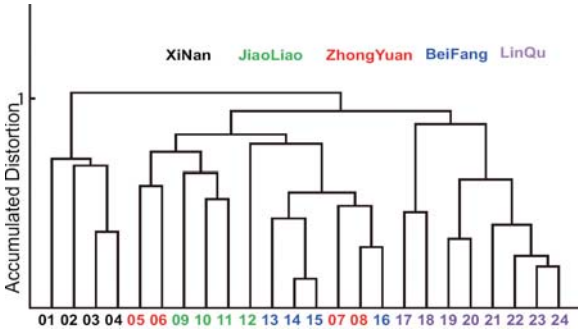

Fig. 6 Classification of speakers based on sub-dialects

4.3. Speaker classification based on sub-dialects

In this paper, using the recorded data, experiments of speaker classification based on their sub-dialects of Mandarin was carried out and the result with the new data of 24 sub-dialects of Mandarin speakers is shown in Fig. 6. 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 MandarBin, are grouped together in a sub-tree. The speakers 09-12 and 13-16, who are from JiaoLiao and BeiFang sub-dialect regions, are mainly clustered to two sub-trees respectively. Speakers 17-24, two adults and 6 children from the same village, are also grouped to a sub-tree. But for the speakers from ZhongYuan sub-dialect region, although speakers 05-06 from YuZhou and speakers 07-08 from ShangQiu are still grouped together individually, the two speaker groups are located apart from each other. Speakers 05-06 are clustered near to the JiaoLiao sub-dialect region and speakers 07-08 are clustered near to the BeiFang 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 [9]. According to [8] and [9], 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. Therefore, we believe that if the initial and tone information is considered together, the results would be more valid linguistically. In conclusion, this result proves that dialect pronunciation structure can work well on extracting the linguistic information of sub-dialects of Mandarin.

5. Conclusions

In this paper, a structural representation of pronunciation, which is inspirited by infants' language acquisition and originally proposed to remove extra-linguistic features from speech, is applied to represent the pronunciation of Chinese dialects. Firstly, this approach is testified in pronunciation assessment by classifying speakers based on their dialects and satisfactory classification result with high robustness to speaker variability is obtained. Meanwhile, this approach is also applied to classifying speakers of Mandarin sub-dialects after a special corpus of sub-dialects of Mandarin is built. This result also shows that these speakers can be linguistically classified with little influence of the age and gender.

Reference

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

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

[3] P. W. Jusczyk, The discovery of spoken language, Bradford Books, 1997.

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

[5] Y. Qiao et al., f-divergence is a generalized invariant measure between distributions. InterSpeech, 2008. 1349-1352.

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

[7] D. Saito et al., Structure to speech – speech generation based on infant like vocal imitation-. InterSpeech, 2008, 1837-1840.

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

[9] J. Hou 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, 2005. vol. 13, no. 5, 930-944.

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