Chinese Dialect Classification Using Acoustic Universal Structure in Speech *

◎ Xuebin Ma, Max Takazawa, Nobuaki Minematsu, Keikichi Hirose (The Univ. of Tokyo)

1 Introduction

It is widely accepted that the performance of modern Automatic Speech Recognition (ASR) systems is strongly affected by the variations between training data and test data. Among these variations, some are caused by non-linguistic factors, such as the physical features of the speakers, the different recording equipments, environments and so on. Meanwhile, some of the variations are also caused by some linguistic factors, for example, the accent and dialect of languages, which always exhibit some regional regularity. Nowadays, in order to deal with these variations to enhance the performance of the ASR systems, a great number of frameworks and algorithms are proposed and some satisfying results are already obtained. But in some particular cases, such as the dialect classification, none of these methods is suitable because they can't distinguish well the linguistic variations from the variations caused by non-linguistic factors.

In China, there are hundreds of dialects and sub-dialects which are more or less different to each other lexically, phonologically and phonetically. Then the classification of these dialects becomes a challenge especially considering that the variations caused by dialectal factors are acoustically mixed with the non-linguistic variations.

In this paper, after introducing the special dialectal situation in China, a new method is proposed to classify dialect speakers by their acoustic structures which are build on their utterances of 9 selected monophthong characters. Then, classification experiment is carried out and the result is discussed.

2 Chinese Dialects and Their Interrelations

2.1 Complex Dialect Situation in China

There are mainly 7 big dialect regions (Guan, Wu, Yue, Min, Xiang, Hakka, Gan) in China. Almost every dialect region has many sub-dialects and these sub-dialects change from one city to another.

Table 1 Sub-dialects and their distribution in China

Dialects	Sub-dialects	Populations	Cities
Guan	42	836	1517
Wu	13	77	153
Yue	8	71	99
Min	9	60	142
Xiang	3	36	71
Hakka	8	34	207
Gan	11	31	102

Sometimes, even between two adjacent cities, the dialect changes and the people have difficulty in understanding the dialect of each other. This complex dialect situation can be demonstrated by Table 1, which shows the number of the sub-dialects in these big dialect regions together with their distribution by population and the number of cities [1].

Since 1956, Mandarin, the main branch of Guan dialect region, has been prescribed all over the country as the official language and almost every Chinese began to learn Mandarin. But, the dialects, especially several big dialects, are still used widely. And even not in their native dialect regions, the people from the same dialect region always like to speak their dialect to each other to show the special close relation.

2.2 Interrelations among Chinese Dialects

In fact, although Chinese dialects are mutually unintelligible to some degrees, they share a lot of common features which are INHERITED from Middle Chinese. For example, they use the same written characters, every character is pronounced as a mono-syllable, every syllable has the same phonological structure and so on. In this paper, as we are currently focusing on the acoustic features, only the phonological features of the dialects and their phonetic realizations are introduced.

By traditional Chinese phonology, Chinese syllables can be divided into two parts, the Initials and

* 音声の構造的表象に基づく中国語方言の分類, 馬学彬, 高澤真章, 峯松信明, 広瀬啓吉 (東京大学)

the Finals. Every Initial is usually consisted of a consonant or a nasal. Every Final is always consisted of a vowel with optional consonantal onset or coda. In Mandarin, there are 22 Initials and 38 Finals including the null units [2]. But in the dialects, the Initial/Final inventory always changes. For example, there are 19 Initials and 53 Finals in Cantonese which is the main branch of Yue dialect region. And in Shanghainese of Wu dialect region, there are 35 Initials and 32 Finals. In addition, even to the same Initial/Final unit in different dialects, the phonetic features change sometimes. So how to classify the dialects just using the acoustic features is still a problem.

3 Acoustic Representation of Dialects

In our previous works, a new speech representation based on the acoustically invariant properties of speech [3] was proposed. It can extract the purely linguistic information in speech. In this paper, we will use this method to extract the purely dialectal information from utterances of dialect speakers and classify the dialect speakers.

3.1 Modeling the Non-linguistic Variations

In speech recognition studies, the spectral variations caused by non-linguistic factors could be classified into three kinds: additive, convolutional and linear transformational distortions. The additive distortion is often caused by background noise. It is not inevitable and we can move to a quiet room to avoid it, so it is ignored in this paper. The convolutional distortions are typically caused by microphones and the reflection in a recording rooms. At last, the vocal tract length difference among speakers is a typical reason of linear transformational distortions. If we represent a speech event by a cepstrum vector c, the convolutional distortion changes c into $c' = c+b$. The linear transformational distortion, which is often modeled as frequency warping, changes c into $c' = Ac$. Then the total distortion of cepstrum c becomes $c' = Ac + b$. These two kinds of spectral distortions can be schematized by Fig. 1. The horizontal and vertical distortions correspond to the distortions due to matrix A and vector b, respectively.

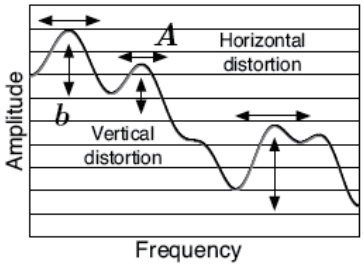

Fig. 1 Spectral distortions caused by matix A and vector b

3.2 Speaker Invariant Structure in Dialects

In this paper, every speech event is captured as a distribution. And all the event-to-event distances can be calculated as Bhattacharyya Distance (BD). So the distances among speech events can be represented by a BD-based distance matrix.

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

Furthermore, by our previous studies, the distance matrix attained by BD was proved mathematically to be invariant to affine transformations such as $c' = Ac+b$. It means the BD-based distance matrix is invariant with the non-linguistic factors. So, using this distance matrix calculated by utterances of dialect speakers, we can build the speaker invariant structure in dialects.

3.3 Comparable Structure among Dialects

Nowadays, some linguistics are focusing on comparing different utterances of the same Initial/Final unit in Mandarin and dialects. And to every Initial/Final unit in Mandarin, its corresponding ones in individual dialects are already listed [4]. Considering these results, we fix a list of monophthong characters which are widely used in Mandarin and dialects to aim at building speaker invariant structures for individual dialect speakers using the distance matrix of his utterances of these characters. After that, the dialect speakers can be classified based on their structures.

In this paper, 9 monophthong characters are selected to build the structure. Table 2 shows a list of the characters and their corresponding syllables together with their monophthong Finals in Mandarin. Then with the acoustic distances of these utterances in dialects, the speaker-invariant structure is built. After that, we can classify the dialects based on the

Table 2 The selected monophthong characters

characters	阿, 比, 波, 痴, 雌, 俄, 耳, 五, 魚
syllables	/a/,/bi/,/bo/,/chi/, /ci/,/e/,/er/,/u/,/ü/
monophthongs	/a/,/i/,/o/,/i/, /i/,/e/,/er/,/u/,/ü/

Table 3 Detailed information of the speakers

Dialects	Cities	Speakers	City ID
Wu	ShangHai	4	1
	SuZhou	1	2
	YiXing	1	3
	ShaoXing	1	4
	NingBo	1	5
Yue	HongKong	2	1
	FoShan	1	2
	MeiXian	1	3
Min	ZhangZhou	1	1
	JiJiang	1	2
	FuJian	1	3
Hakka	MeXian	2	1

structures.

4 Chinese Dialect Classification Experiment

4.1 Speech Data Used in the Experiment

In order to build the monophthong structures to classify the dialect speakers, we recorded some data in the University of Tokyo. Seventeen Chinese graduate students joined our recording. All the students were native dialect speakers. All the speakers were asked to read the selected characters in their native dialects. Every character was read four times and all the data were recorded in a sound proof room. The detailed information about the speakers is shown in Table 3.

4.2 Calculation of the Structures

After the data of every dialect speaker was recored, the utterances of the selected monophthong characters were analyzed under the acoustic conditions in Table 4. Then the BD of any two utterances was calculated and the speaker invariant structure was built.

Table 4 Acoustic conditions of the analysis

sampling	16bit / 16KHz
windows	Blackman window 25ms length,1ms shift
parameters	Mel-cepstrum (1-10 dimensions)
distribution	Gaussian distribution estimated with MAP

By checking the built structure of every speaker, the speaker invariance can be proved experimentally. The tree diagrams of Fig. 2 show the utterance structures of three Cantonese speakers which are drawn with Ward's method. One of the speaker is female, the other two are male. The first two speakers are from Hong Kong and the third speaker is from GuangZhou. By this figure, we can find that the structures of first two speakers are almost the same although the size of two structures is a little different to each other and the third speaker's structure is structurally similar to the others. So it is proved that the linguistic information of their utterances is purely extracted.

4.3 Measurement of the Structure Distances

The distance between two structures is obtained by shifting($+b$) and rotating($\times A$) one structure until the best overlap between them is observed just like Fig. 3. It means the minimum of the total distance between the corresponding two points of the two structures. In [5], it was already proved that the minimum total distance can be approximately calculated as Euclidean distance between two distances matrices. The formulation is shown, where follows while the S_{ij} means (i, j) element of matrix S and M means the number of the utterances (monophthongs in this paper).

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

4.4 Classification Results and Discussions

Fig. 4 shows the result attained by classifying all the dialect speakers by their distance structures. The letters indicate the big dialect region of the speakers and the numbers mean which city the

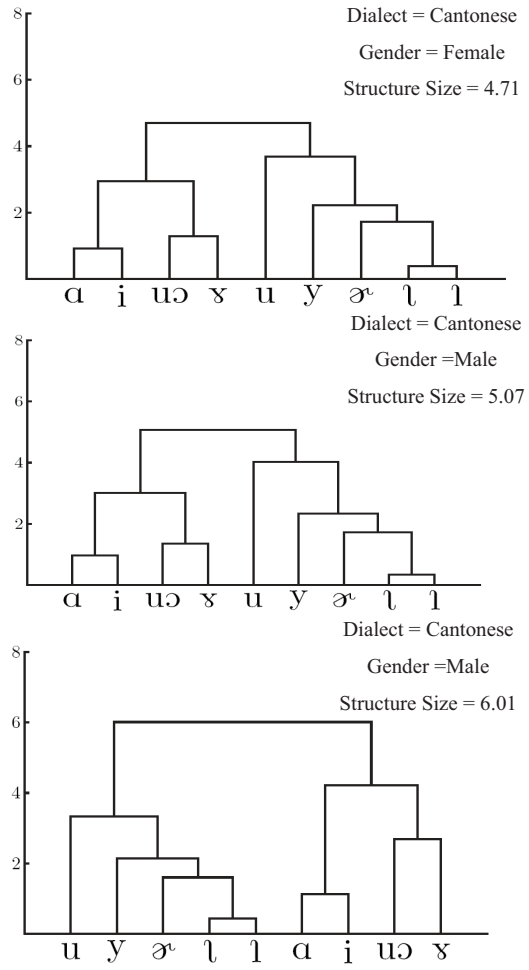

Fig. 2 The structure of three Cantonese speakers

speaker is from. By the result, we can see the speakers are completely classified by their dialects. The relations among the sub-dialects are also well shown as the speakers from the same city are plotted near to each other too. There is also an exception, the structure of a speaker who is from city 5 of Wu dialects is found very near to the Hakka speaker. Anyway, considering that the structures are built just using the 9 monphthongs and the city 5 (Ningbo) is very near to the Hakka region geographically, the result is quite satisfactory.

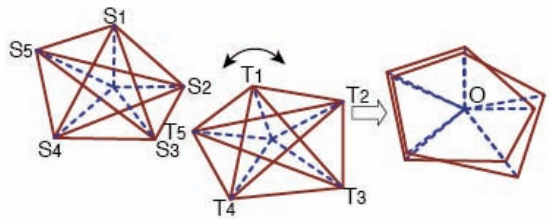

Fig. 3 Distance calculation after shift and rotation

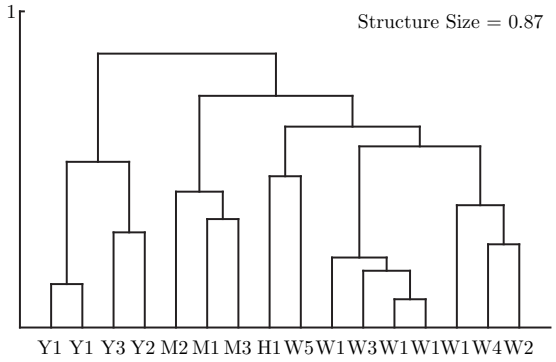

Fig. 4 Classification of the dialect speakers

5 Conclusions and Future Work

This paper proposed a new method to classify the dialect speakers by building the speaker-invariant utterance structures. And satisfactory results were obtained in our dialect classification experiment. In the near future, we are going to classify more speakers utilizing some linguistic knowledge and do some accent classification if some Chinese accent databases are available.

References

[1] http://www.glossika.com/en/dict/

[2] Lin-shan Lee, "Structural Features of Chinese Language – Why Chinese Spoken Language Processing is Special and Where We Are, keynote Speech", International Symposium on Chinese Spoken Language Processing, Singapore, pp.1-15 (1998)

[3] N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for CALL", Int. Workshop on Spoken Language Technology (SLT'2006), pp.126-129 (2006-12)

[4] Yuan Jiahua et al, HanYu FangYan GaiYao, YuWen ChuBanShe, YuWen Press (2000)

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