Dialect-based Speaker Classification of Chinese Using Structural Representation of Pronunciation

Xuebin MA[1], *Nobuaki MINEMATSU*[2], *Yu QIAO*[2], *Keikichi HIROSE*[3], *Akira NEMOTO*[4], *Feng SHI*[4]

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

Abstract

In China, there are thousand kinds of dialects and sub-dialects. Because there are many differences among them in varying degrees, grammatically, lexically, phonologically and phonetically, people from different dialect regions always have some difficulties in oral communication. In addition, many of these dialects are still developing and their linguistic features are also changing under the influence of standard Mandarin, which has been popularized all over the country by the government. In this paper, Chinese dialect-based speaker classification using speech technologies is discussed. Since acoustic features of utterances convey not only linguistic information such as dialectal information, but also extra-linguistic information such as age, gender, speaker, etc, we have to focus on only the dialectal features in speech. For that purpose, to classify Chinese speakers, we use the structural representation of pronunciation, which was originally proposed to remove extra-linguistic features from speech [1, 2]. We carry out some experiments after we built a special corpus composed of dialectal utterances of selected characters. This is because the publicly available Chinese dialect corpora were developed for different purposes and can hardly be used in this study. Using the utterances of a special set of Chinese characters, each speaker is modeled as his/her pronunciation structure. Then, all the structures are classified based on bottom-up clustering. The proposed method is also tested especially in terms of robustness to speaker variability. Here, the utterances of simulated very tall and short speakers are classified. Further, a comparative experiment using the conventional method is also carried out. All the results show that our approach can perform highly linguistically-reasonable classification.

1. Introduction

In modern speech technologies, spectrum is often used to represent the segmental aspect of speech. However, it carries not only linguistic information but also extra-linguistic information corresponding to age, gender, speaker, and so on. In other words, the same linguistic content is acoustically realized differently from a speaker to another. If one wants to classify speakers based on their dialects, one has to focus only on the acoustic features which are relevant to dialectal differences. This is because two utterances of the same linguistic content spoken by a very tall adult and a very short child are more different acoustically than an utterance of an adult and its dialectal version of that adult. In the case of automatic speech recognition (ASR), a similar problem happens because the aim of ASR is to extract lexical information from an utterance by ignoring speaker information. For this aim, speaker-independent acoustic models are often built by collecting utterances from thousands of speakers. Logically speaking, however, in dialect-based speaker classification, this approach cannot be accepted. The aim of the latter is classifying speakers only based on their dialects and this aim cannot be attained with dialect models trained with utterances from different speakers of the individual dialects. If one is interested in dialect identification, the dialect or accent models can be built with different speakers [3, 4], which were also used in accent analysis and evaluation [5, 6, 7]. However, if one wants to focus on intra-dialect relations among speakers of that dialect, it is not desired to create a dialect model using different speakers of the same dialect. Considering the current complicated situation of Chinese dialects, we can say that speakers of the same dialect are often speakers of different sub-dialects.

In our previous work, a structural representation of speech was proposed to remove extra-linguistic and irrelevant acoustic features from utterances [1, 2]. This speech structure is calculated by extracting speaker-invariant speech contrasts or dynamics and it shows high speaker independence. Using speech structures, speaker-independent ASR was realized only with a small number of training speakers, where explicit adaptation or normalization was not needed [8, 9]. Further, the structures were already applied for CALL [10] and speech synthesis [11] and satisfactory results were obtained.

This paper describes the first trial of applying the speech structures to Chinese dialect-based speaker classification. In Section 2, some fundamentals of Chinese dialects are introduced. After briefly describing speaker-invariant but dialect-sensitive acoustic structures in Section 3, the effectiveness of the speech structure in Chinese dialect-based speaker classification is examined in Section 4. Further, in Section 5, by using utterances of simulated children, we also examine the proposed method especially in terms of the robustness to speaker variability. This paper is concluded and the future works are described in Section 6.

2. Fundamentals of Chinese dialects

In China, there are mainly 7 big dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue, Min) [12] and most of them have some different sub-dialects and sub-sub-dialects too. For example, dialects in Guanhua (Mandarin) region can be further grouped

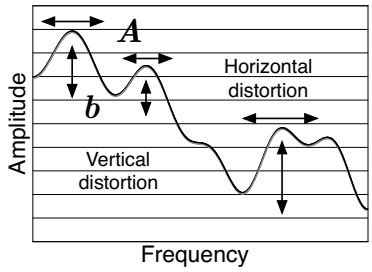


Figure 1: Spectral distortions caused by matix *A* and vector *b*

into 7 sub-dialects and 42 sub-sub-dialects [13]. Nevertheless, all these dialects and sub-dialects are developed from Middle Chinese, which is referred to as the Chinese spoken language during the period from 6th to 10th century, and a lot of common features have been 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 as 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. However, among these dialects, 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 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 regional accent, if he/she has some knowledge of these dialects. On the other hand, as standard Mandarin are 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 by the influence of Mandarin or other dialects. 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 often show somewhat different linguistic features because they may belong to different sub-dialects. So in dialect-based speaker classification, it is necessary to consider the dialectal features of individual speakers through removing extra-linguistic features.

3. Structural representation of dialects

3.1. Modeling extra-linguistic information mathematically

After utterances are 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 microphone differences, and vocal tract length differences are the typical reason of linear transformational distortions [14]. 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, where the horizontal and vertical distortions correspond to the distortions due to matrix A and vector b, respectively.

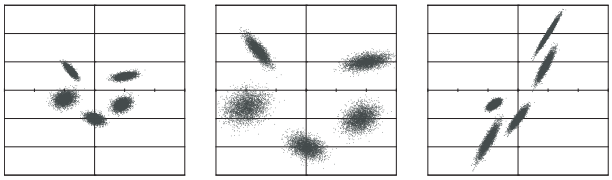


Figure 2: The invariant underlying structure among three data sets

3.2. Speaker-invariant structure in dialects

As non-linguistic variation in speech is modeled as affine-transform, to obtain speech features invariant to non-linguistic variation, we have to use affine-invariant features. In [9], Bhattacharyya Distance is shown to be invariant with affine transform. Here, every speech event is captured as a distribution ($p_i(c)$) and event-to-event distances are calculated as Bhattacharyya Distance (BD).

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

By calculating BDs between any pair of speech events, a distance matrix can be obtained. Since a distance matrix can represent uniquely its geometrical shape composed of all the speech events, we call the matrix a pronunciation structure of these speech events. With the utterances of dialect speakers, we can build structural representations of the dialect speakers which are invariant to extra-linguistic factors. Fig. 2 shows an example of that invariant underlying structure among three sets of speech events. Any set of the events are obtained by affine transform of either of the other two sets. This means that the BD-based distance matrix is invariant and common among the three sets. 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.

3.3. Building comparable dialectal structures

In order to classify speakers based on their dialects using structural representation, comparable dialectal structures should be built from their dialectal utterances containing the same set of some linguistic units. Considering there are many grammatical and lexical differences among Chinese dialects, syllable or smaller phonological units can be a good choice as the linguistic unit. However, although all Chinese dialects are sharing the same phonological structures, the inventory of their phonological units of Chinese dialects are different. Considering that

Table 1: Selected characters and their pronunciation in Mandarin

Characters	笔,思,十,耳,五,鱼,阿, 波,饿,哀,悲,早,肉,左, 鸭,哇,别,月,歪,对,腰,牛, 安,烟,弯,捐,恩,彬,温,君, 央,帮,汪,崩,冰,翁,宗,用
Mandarin Pronunciation	/bi/,/ci/,/shi/,/er/,/wu/,/yü/,/a/,/bo/,/e/, /ai/,/bei/,/zao/,/rou/,/zuo/,/ya/,/wa/,/bie/, /yue/,/uai/,/dui/,/yao/,/niu/,/an/,/yan/,/wan/, /juan/,/en/,/bin/,/wen/,/jun/,/ang/,/yang/, /wang/,/beng/,/bing/,/weng/,/zong/,/yong/

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 pronunciation structure for dialect comparison. If we can select a common list of characters which can cover most of the phonological units in all the dialects, reasonable and comparable structures for the dialects 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. For example, using the dialectal utterances of the same written characters, initial/final units of different dialects are listed and their phonetic features are compared. As a result, the relations of these units between Mandarin and other dialects are often shown. In [12], all the initial/final units in different dialects and their corresponding ones in Mandarin are listed together with some characters as examples. Based on these studies, we fixed a common list of 38 characters which are covering the 38 finals in Mandarin. These characters and their corresponding pronunciations in Mandarin are listed in Table 1. The dialectal pronunciation structure for every dialect speaker can be built from his/her dialectal utterances of the selected characters and it is invariant with extra-linguistic factors. Then the speakers can be classified based on their dialects using these comparable dialectal structures.

Nowadays, speech corpora are regarded as the most important infrastructure of modern spoken language technologies and many Mandarin corpora have been built. As for dialect speech corpora, only several ones are available to the public but most of them cover an individual dialect. Besides, these corpora are developed for different purposes and can hardly be combined and used in this study. Further, most available corpora are developed for ASR using utterances in conversation but controlled utterances are needed in this study. So using Table 1, we recorded some dialect data of Chinese speakers from different cities for the experiments.

4. Classification experiments

4.1. Speech materials used in the experiment

For preliminary experiments, we recorded some data of 17 Chinese dialect speakers. They are all native speakers and most of them were born and brought up all the time in the same dialect regions, except one female speaker. Her parents are both native Hakka speakers and they moved to a Cantonese region when she was 10 years old. So she has mastered two dialects, Hakka and Cantonese. All the subjects keep speaking their dialects although living in Japan, at least during the conversation with their families and friends from the same dialect regions. In the experiment, every speaker was given a speaker ID and details of their hometown and gender are listed in Table 2, where their dialect regions are represented by color and the ID of female speakers are represented by italic type. The above mentioned female speaker has two speaker IDs, 02 and 07, which stand for her two dialects, Hakka and Cantonese, respectively. All the data were recorded in a sound proof room. The speakers were asked to read the selected characters of Table 1 in their native dialects, and the dialectal pronunciation of these selected characters were all checked before the recoding. During the recording, each character was read four times. After that, every syllable was detected manually and stored into individual files. Then, these data were analyzed under the acoustic conditions shown in Table 3. Each speech event (syllable or vowel) was modeled as diagonal Gaussian distribution and the parameter estimation was done for Gaussian modeling using MAP (Maximum A Posteriori) criterion.

Table 2: Detailed information of the speakers

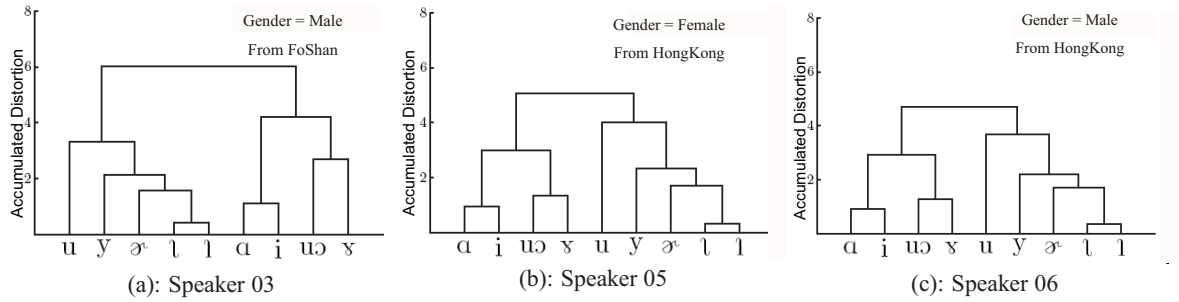
Speaker ID	Dialect	Cities	Gender
01	Kejia	DaBo	M
02	Kejia	ShenZhen	F
03	Yue	FoShan	M
04	Yue	MeiXian	F
05	Yue	HongKong	M
06	Yue	HongKong	F
07	Yue	ShenZhen	F
08	Min	ZhangZhou	M
09	Min	FuZhou	F
10	Min	JiJiang	M
11	Wu	ShangHai	M
12	Wu	ShangHai	M
13	Wu	ShangHai	M
14	Wu	ShangHai	F
15	Wu	ShaoXing	M
16	Wu	NingBo	M
17	Wu	YiXing	M
18	Wu	SuZhou	F

Table 3: Acoustic analysis condition

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

4.2. Phonetic tree of monophthongs

In Mandarin, there are 9 monophthongs and they are covered by the first 9 selected characters of Table 1. By using the final parts of the utterances of these characters, the monophthong structure (distance matrix) for each speaker can be built. Firstly, the final parts of these utterances are detected manually and each of them is modeled as a single Gaussian individually. Then the BDs of every pair of monophthongs are calculated for each speaker to



(a): Speaker 03 (b): Speaker 05 (c): Speaker 06

Figure 3: Phonetic tree of three Cantonese speakers

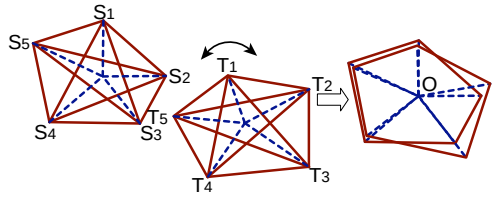


Figure 4: Distance calculation after shift and rotation

form his/her structure. A distance matrix can be visualized as a tree diagram by Ward's clustering method. Fig. 3 shows the trees of three Cantonese speakers of 03, 05 and 06. Speaker 03 is male from FoShan, speaker 05 is male from HongKong, and speaker 06 is female from HongKong, too. The nodes are the IPA symbols of the 9 monophthongs in Mandarin. We can see the phonetic trees of speaker 03 and 05 are structurally very similar but slightly different. Locally speaking, a difference between ɤ and uɔ in 03 is larger than in 05. Globally speaking, they are very similar considering that the mirrored position of the two sub-trees can be ignored in tree diagrams. Meanwhile, we can see the phonetic trees of speaker 05 and 06 are almost the same although they are different gender. The result shows the phonetic tree of a speaker, the pronunciation structure of a dialectal speaker, is sensitive to dialectal information and highly independent of genders.

4.3. Distances between syllable-based structures

In Fig. 3, the structures are obtained by monophthongs, which are calculated using the final part of dialect syllables. In fact, more dialectal features can be found by syllable-based analysis, where syllable-to-syllable distances have to be calculated. There are two methods to calculate this distance. One method is that a whole syllable is modeled as a Gaussian, just as in building the monophthong structures, each monophthong segment was modeled as Gaussian. Another is that a syllable is modeled as a sequence of a fixed number of distributions, such as HMM. Syllable-to-syllable distance is obtained as summation of distances between the corresponding distributions. Since these Chinese syllables are all very short, we adopted the first method.

By calculating the BD of every pair of syllables, a 38×38 distance matrix can be obtained by calculating the BD of every pair of syllables, which fixes the unique pronunciation structure for that speaker. Then, the distance between two structures is obtained after one is shifted ($+b$) and rotated ($\times A$) [16] until

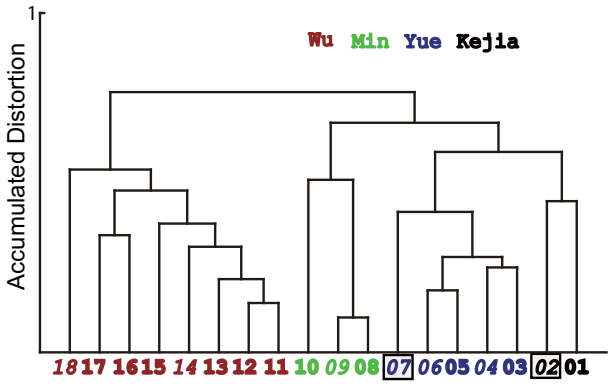


Figure 5: Classification of the dialect speakers

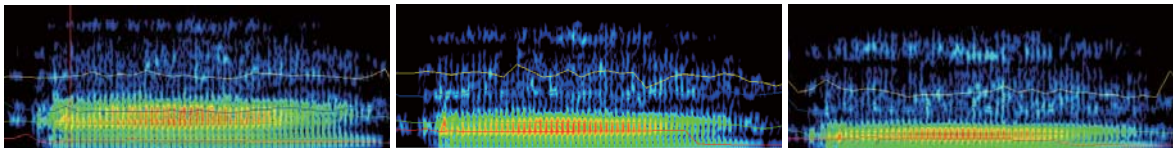
the best overlap is observed between them, which is shown in Fig. 4. With the best overlap after shift and rotation, the distance between two structures is calculated as the minimum sum of the distances between the corresponding two events of the two structures. In [1], it was experimentally proved that the minimum sum can be approximately calculated as Euclidean distance between two distance matrices. Following is the 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 of speakers A and B, respectively. M means the number of the syllables. Although the vowel trees in Fig. 3 were obtained from manually segmented vowel segments, the following experimental results are obtained by an automatic procedure.

4.4. Classification result and discussions

Using inter-speaker distances of D_1, the dialect speakers are classified using Ward's clustering method and the result is shown in Fig. 5, where every speaker is represented by speaker ID in Table 2. The dialect regions are shown by colors used in Table 2. In the figure, the speakers from the same dialect region are clustered together. Besides, the speakers from the same sub-dialect are also clustered nearer to each other. For example, speakers 11-14 are all from the same city of Wu dialect region and they are classified into a sub-group in the result. Meanwhile, we can see that the speakers from Min, Yue and Kejia dialects regions are clustered into a big group. It can be explained



(a): Short speaker (b): Original speaker (c): Tall speaker

Figure 6: Spectrums of short and tall speakers

not only by the reason that these dialect regions are very close to each other geographically, but also because it is proved by some historical linguists that these dialect regions are affected by each other greatly during their development [12, 15]. The result also shows high independence of the gender of the speakers and other extra-linguistic factors. For example, as described in Section 4.1, 02 and 07 are the same speaker with different dialects and they are classified into their corresponding dialect groups correctly, not be classified into the same group. In conclusion, this result shows that these speakers are classified only by the purely linguistic information of their utterances.

5. Experiments with simulated data

5.1. Simulated data of tall and short speakers

It is known that the vocal tract length of speakers is an important extra-linguistic factor and which is generally determined by the height of speakers, a tall speaker has a long vocal tract and a short speaker has a short vocal tract, generally. We can use a frequency warping function and simulate the utterances of speakers as if they are produced by the same speaker of much longer or shorter vocal tract. Frequency warping is characterized in the cepstral domain by multiplying c by matrix A ($=\{a_{ij}\}$) [14].

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

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. Otherwise, when $\alpha > 0$, formants are transformed to be higher and the vocal tract length shorter. Using matrix A, the recorded data were converted into a shorter version with $\alpha = 0.2$ and a taller version with $\alpha = -0.2$ using STRAIGHT [17]. Fig. 6 shows the spectrums of the same syllable produced by a Cantonese speaker and his two simulated versions. From left to right is the pseudo short speaker, original speaker and tall speaker.

5.2. Experimental result and discussions

Using the original and simulated data together, we did the same classification experiments and the result is Fig. 7. In the figure, the original speakers are represented by the speaker IDs as in Table 2, the same ID with a line on the top represents the tall

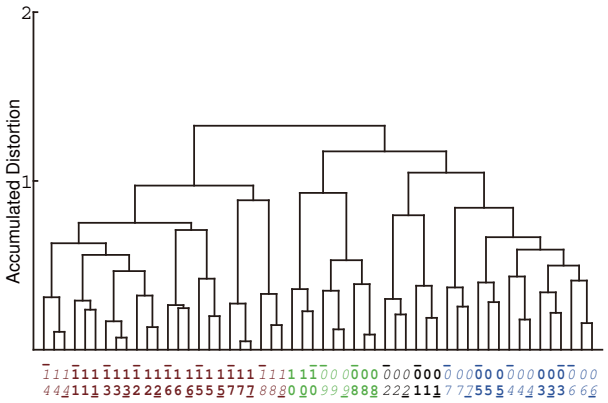


Figure 7: Classification of original, tall and short speakers using D_1

version of this speaker and the same ID with a line on the bottom represents the short version. Then, in Fig. 7, we can see the structure is really speaker invariant because the original speaker and the two simulated shorter and taller speakers are all clustered together. Similar to Fig. 5, the speakers from the same dialect region or sub-dialect region are clustered together and the speakers form Min, Yue, Kejia dialects regions are clustered into a big group. This proves that the proposed method does work well on extracting the purely dialectal features and really speaker-invariant.

5.3. Result of a comparative experiment

In conventional acoustic framework such as DTW and HMM, speech events of a speaker are directly compared acoustically with those of another speaker and, in this framework, the distance between two dialectal syllable structures is formulated as

$$D_2(A,B) = \sqrt{\frac{1}{M} \sum_i BD(S_i^A, S_i^B)}. \quad (4)$$

S_i^A is syllable i of speaker A and S_i^B is syllable i of speaker B. M means the number of the syllables. In [3, 4], the phonemes of a language or dialect are acoustically modeled as distributions (GMM) or sequences of distributions (HMM) by collecting a large number of speakers of that language or dialect. In these works, distributions of cepstrums (spectrums) are used. If these methods are applied directly for speaker classification, basically speaking, they come to compare two speakers based on D_2. Then using D_2 for the same simulated data set, the classification result is obtained and shown in Fig. 8. Similar to Fig. 7, the original speakers are represented by the speaker IDs, the same ID with a line on the top represents the tall version of this

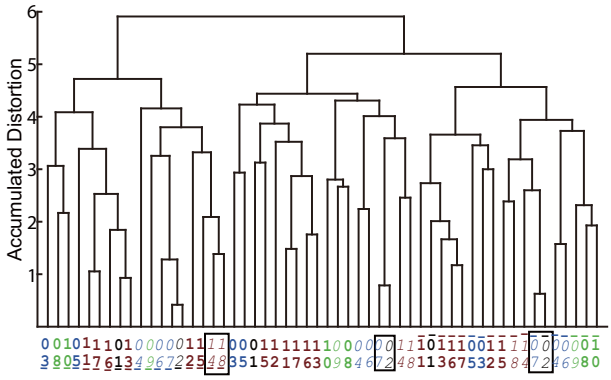


Figure 8: Classification of original, tall and short speakers using D_2

speaker and the same ID with a line on the bottom represents the short version. Although using the same data set that was used in Fig. 7, by contrast, the speakers are classified into three big sub-trees corresponding to their vocal tract length. The simulated short speakers are all classified into the left sub-tree, the simulated tall speakers are all classified into the right sub-tree and the original speakers are classified into the middle sub-tree. And in every sub-tree, by checking the position of the IDs in italic type which means the female speakers, it can be found that the speakers of the same genders are mainly clustered together. Meanwhile, the speakers are clustered with no relation to their dialects. In each sub-tree, 07 and 02, who are the same speaker speaking different dialects, are judged to be very close to each other, just as we expected in section 3.2. Comparing to Fig. 7, these results proved again that our proposed method does work well on extracting the speaker-invariant dialectal features.

6. Conclusions and future works

In this paper, we proposed to use the structural representation of pronunciation to classify dialect speakers. We first selected a list of written characters considering the phonological features of Chinese dialects, and built the pronunciation structure for each speaker using his/her dialectal utterances of these characters. After that, based on the distances among these structures, dialect-based speaker classification experiments are conducted and a satisfactory result is obtained. Finally, using the utterances of simulated tall and short speakers obtained by frequency warping in the cepstral domain, the method is also tested in classification experiment. And comparing to conventional method of calculating the distances between syllable structures, a comparative experiment is carried out. All the results show this method performs very good linguistically-reasonable classifications and the structural representation is highly independent of extra-linguistic variations caused by speaker variability.

Currently, a larger corpus of more dialect speakers are being developed in the mainland of China. With this larger corpus, some other results will be presented in the conference. Meanwhile, considering the current situation that many people are speaking Mandarin with regional accents, some data of Mandarin with regional accents are also recorded and this proposed method will be applied in speaker classifications based on their accented Mandarin. As future works, some applications using this approach will be developed, such as detecting the dialectal information of the speakers, testing the acoustic distances among Chinese dialects and evaluating the Mandarin pronunciation in Computer Aided Language Learning (CALL).

7. 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] M.A. Zissman et al., "Automatic language identification," Speech Communication, vol. 35, no. 1-2, pp. 115-124, 2001.

[4] W.H. Tsai et al., "Discriminative training of Gaussian mixture bigram models with application to Chinese dialect identification," Speech Communication, vol. 36, no. 3-4, pp. 317-326, 2002.

[5] S.A. Ghorshi et al., "Cross entropy information metric for quantification and cluster analysis of accents," Int. Workshop on Speech Recognition and Intrinsic Variations, pp. 119-122, 2006.

[6] M. Huckvale, "Accdist: A metric for comparing speakers' accents," ICSLP, pp. 29-32, 2004.

[7] S. Wei et al., "Automatic mandarin pronunciation scoring for native learners with dialect accent," ICSLP, pp. 1383-1386, 2006.

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

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

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

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

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

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

[14] 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.

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

[16] D. Saito et al., "Decomposition of rotational distortion caused by VTL difference using eigenvalues of its transofmation matrix," INTERSPEECH, pp. 1361-1364, 2008.

[17] H. Kawahara et al., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneousfrequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187-207, 1999.