社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

非言語的な要因に不変な音響的特徴を用いた中国語方言に基づく話者分類

馬 学彬† 峯松 信明† 喬 宇† 広瀬 啓吉† 根本 晃†† 石 峰††

† 日本東京大学 〒 113–8656 日本東京都文京区本郷 7–3–1
†† 中国南開大学 〒 300071 中国天津南開区衛津路 94 号
E-mail: †{xuebin,mine,qiao,hirose}@gavo.t.u-tokyo.ac.jp, ††akiranmt@hotmail.com,
†††shifeng@nankai.edu.cn

あらまし 現行の音声技術を利用した中国語方言に基づく話者の分類は、中国語方言自身の持つ複雑さや、方言音の音響的特徴が方言的情報だけでなく性差、年齢差、個人差などの非言語的情報をも伝達することから容易ではない。本研究では、音声から非言語的情報を除去した発音の構造的表象を用いた方言話者分類法を提案する。この手法では、漢字音セットに基づく方言音の録音を経て、各話者は各人の発音構造としてモデリングされ、多数話者の方言的構造群はボトムアップクラスタリングにより分類される。本稿では、話者の多様性に対するロバスト性についても同様に検証する。ここでは我々が人工的に合成した巨人と子供の音声を用い、従来手法との比較実験を行う。実験の結果、中国語方言の事情にも合致し、また年齢、性別に対し高い独立性のある結果が得られた。

キーワード 中国語方言分類、構造的表象、発音、言語的特徴、声道長

Dialect-based speaker classification of Chinese using acoustic features invariant with extra-linguistic factors

Xuebin MA†, Nobuaki MINEMATSU†, Yu QIAO†, Keikichi HIROSE†, Akira NEMOTO††, and Feng SHI††

† Univ. of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo 113–8656 Japan
†† Nankai Univ., 94 Wenjin Rd, Nankai, Tianjin, 300071 P.R.China
E-mail: †{xuebin,mine,qiao,hirose}@gavo.t.u-tokyo.ac.jp, ††akiranmt@hotmail.com,
†††shifeng@nankai.edu.cn

Abstract Chinese dialects-based speaker classification using modern speech technologies is really a challenge, not only because the situation of Chinese dialects is very complicated, but also because acoustic features of utterances convey dialectal information together with extra-linguistic information such as age, gender, speaker, etc. In this paper, we propose a new speaker classification technique using structural representation of pronunciation, which was originally proposed to remove extra-linguistic information from speech. After collecting dialectal utterances of a selected set of Chinese characters, each speaker is modeled as his/her pronunciation structure. Then, all the dialectal structures are classified based on bottom-up clustering. We also test the proposed method especially in terms of robustness to speaker variability. Here, the utterances of simulated very tall and short speakers are classified to compare the proposed method and the conventional one. All the experimental results show linguistically-reasonable classifications and the high independence of age and gender.

Key words Chinese dialect classification, structural representation, pronunciation, linguistic features, vocal tract length

1. Introduction

In modern speech processing technologies, acoustic features of speech are usually represented by spectrum, which contains not only linguistic information but also extra-linguistic information corresponding to age, gender, speaker, microphone and so on. It means the same linguistic content is acoustically realized differently from a speaker to another. This kind of problem is also the reason why automatic speech recognition (ASR) system trained with a specific group doesn't work well with another group. In fact, the aim of ASR is to extract the linguistic information from an utterance by ignoring speaker information. Trying to solve this problem, speaker-independent acoustic models are often trained by collecting utterances from thousands of speakers but speaker adaptation or normalization techniques are still required too. In dialect or accent identification, the dialect or accent acoustic models are often built with the utterances of many different speakers [1], [2], which were also used in accent analysis or evaluation [3]〜[5]. Nevertheless, this approach cannot be accepted in classifying speakers based on their dialects. In this case, their acoustic features relevant only to dialectal differences are needed. This aim cannot be attained by training individual dialect models with utterances from different speakers. In other words, intra-dialect relations among speakers are needed to be focused on and it is not desired to create a dialect model from utterances of different speakers of the same dialect, because speakers of the same dialect are often speakers of different sub-dialects. There are so many different dialects and sub-dialects in China.

In our previous work, a structural representation of speech was proposed to remove the acoustic features caused by extra-linguistic factors [6], [7]. After modeling the speech variations caused by extra-linguistic factors mathematically, this speech structure is calculated by extracting speaker-invariant speech contrasts and it shows high speaker independence. In our previous study, speech structures were used in ASR system which were trained with a small number of speakers and without explicit speaker adaptation or normalization [8], [9]. It showed much higher performance than widely used methods especially in mismatched conditions. Further, this structural representation was also applied for CALL [10]. By building and showing the vowel structures, language learners can be classified and indicated which vowels should be corrected by priority. Recently, the structures were also applied for speech synthesis and satisfactory results were obtained [11].

In this paper, speech structures are applied for Chinese dialect-based speaker classification by extracting the acoustic features relevant only to dialectal differences. As these features are invariant with extra-linguistic factors, the speakers are classified only based on their dialects. At the beginning, some fundamental knowledges about Chinese dialects and the current complicated situations are introduced in Section 2. In Section 3, it is presented that how the dialectal speech structures are extracted, which are invariant with extra-linguistic factors. In Section 4, some experiments with recorded dialectal data are carried out to examine the effectiveness of this dialect speech structure. In Section 5, by using utterances of simulated very tall and short speakers, the proposed approach is examined especially in terms of the robustness to speaker variability and a comparison experiment is also given using the same data set. At last, the paper is concluded in Section 6.

2. Background of Chinese dialects

Because of many different reasons, the current situation of Chinese dialects is very complicated. In China, there are hundreds kinds of dialects and the main branches of them are classified into 7 big dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue, Min) [12]. Sometimes they are regarded as different languages just like French and Italian, as they are different to each other grammatically, lexically, phonologically and phonetically. So people from these different dialect regions always have difficulty in oral communication. However, all these dialects are developed from Ancient Chinese and Middle Chinese, they inherited a lot of common features. They share the same written scripts, similar sound systems, the same phonological structure and similar phonetic features, etc. For example, every Chinese character is pronounced as mono-syllable with the same structure which is combined by an initial at the beginning, a final at the end and a tone. Further, most of these dialects regions have some sub-dialects and many sub-dialects also have sub-sub-dialects too. For example, GuanHua (also called Mandarin) dialect region has at least 8 big sub-dialects and the sub-dialects of two adjacent cities are different in some phonological or phonetic features. Nevertheless, people from different sub-dialects of the same dialect region can always communicate orally with each other. Since 1956, standard Mandarin has been popularized all over the country as the official language and almost every dialect speaker began to learn Mandarin just like a second language. But being affected by their native dialects, many of them speak Mandarin with some accents. Generally speaking, one can guess the native dialect of the speaker easily according to his/her accented Mandarin, if the hearer has some knowledge about it. On the other hand, as standard Mandarin becomes more and more popular and more and more people move across the country, some dialects are affected and have lost some of their own dialec-

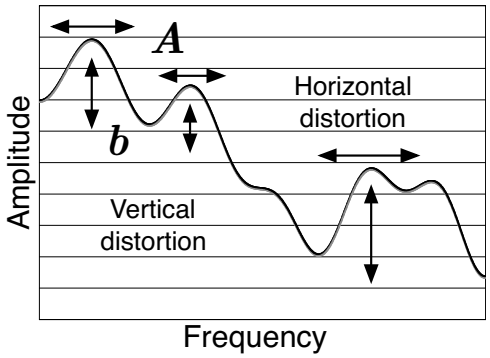

Fig. 1 Spectral distortions caused by matix A and vector b

tal features. However, these dialects, especially some major dialects, are still widely used. People from the same dialect region always like to speak their own dialect to each other to show the special close relationship between them, even outside their native dialect regions.

In brief, dialect-based speaker classification of Chinese becomes more difficult, not only because the linguistic features of Chinese dialects change easily, but also because every speaker has his/her own dialect. Strictly speaking, the pronunciations of two speakers of the same dialect show somewhat different linguistic features. So in dialect-based speaker classification, it is necessary to consider the dialectal features of every speaker, which is invariant with extra-linguistic factors.

3. Comparable structures of dialects

3.1 Acoustic modeling of extra-linguistic features

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 different recording microphones and vocal tract length differences are the typical reason of linear transformational distortions [15]. If a speech event is represented by cepstrum vector c, the convolutional distortion is represented as addition of another vector b and change c into $c' = c + b$. Meanwhile, the linear transformational distortion is modeled as frequency warping of the log spectrum and change 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 while the horizontal and vertical distortions correspond to the distortions due to matrix A and vector b, respectively.

3.2 Structural representation of dialects

As the acoustic features caused by extra-linguistic factors can be modeled as affine transformations, we can build an acoustic structure which is invariant to extra-linguistic factors if this structure is invariant to affine transformations. In fact, every speech event can be captured as a distribution and event-to-event distances are calculated as Bhattacharyya Distance (BD).

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

By calculating BDs between any pair of speech events, a distance matrix can be obtained. Since a distance matrix can fix uniquely its geometrical shape composed of all the speech events, we call the matrix a pronunciation structure of these speech events. The built structure with BD is invariant to extra-linguistic factors as BD is invariant to affine transformation. With the utterances of dialect speakers, we can build structural representations of dialect speakers which are 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 two speakers of different dialects, the difference will be large but it is independent of age and gender.

3.3 Building comparable dialectal structures

In order to classify speakers based on their dialects using structural representation, comparable dialectal structures should be built by using their dialectal utterances of 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. However, although all Chinese dialects are sharing the same phonological units, the inventory of their phonological units of Chinese dialects change. Considering that 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 the individual dialects. If we can select a list of characters which can cover most of the phonological units in the dialects, comparable structures for the dialects can be built with the dialectal utterances of these characters.

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. The relations of these units between Mandarin and other dialects are always given as the results. In [12], all the inital/final units in dialects and their corresponding ones in Mandarin are given together with some characters as examples. Based on their results, we fixed a list of characters which is covering the 38 finals in Mandarin. These characters and their corresponding syllables in Mandarin are listed in Table 1. The dialectal pronunciation structure for every dialect speaker can be built

Table 1 Selected characters in Mandarin

Characters	筆, 思, 十, 耳, 五, 魚, 阿 波, 餓, 哀, 悲, 早, 肉, 左, 鴨, 哇, 別, 月, 歪, 对, 腰, 牛, 安, 烟, 弯, 捐, 恩, 彬, 温, 君, 央, 幇, 汪, 崩, 冰, 翁, 宗, 用
Syllables	/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/

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

with his dialectal utterances of the selected characters and it is invariant with extra-linguistic factors. Then the dialect speakers can be classified based on their dialects using these comparable dialectal structures.

4. Classification of the dialect speakers

4.1 Speech material used in the experiment

17 Chinese graduate students in University of Tokyo joined the recording. They are all native speakers and most of them were born and brought up 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 more details of their hometown and gender are listed in Table 2. The above mentioned female speaker has two speaker

Table 3 Acoustic analysis condition

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

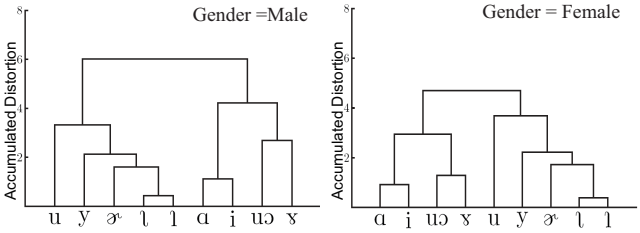

Fig. 2 Phonetic tree of two Cantonese speakers

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 each character was read four times. Then, these data were analyzed under the acoustic conditions shown in Table 3. Each speech event was modeled as diagonal Gaussian distribution and the parameter estimation was done using MAP (Maximum A Posteriori) criterion.

4.2 Phonetic tree of monophthongs

After the structure of every speaker using their utterances of phonemes was calculated, it can be shown as phonetic tree of these phonemes. In Mandarin, there are 9 monophthongs and they are covered by the first 9 selected characters. By using the final parts of the utterances of these characters, the phonetic tree of monophthongs 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 for each speaker are calculated and the monophthong matrix is obtained. At last, the phonetic tree is obtained using a bottom-up clustering method for each speaker. In Fig. 2, the phonetic trees of two Cantonese speakers are shown, speaker 03 and speaker 06. Speaker 03 is a male from FoShan and speaker 06 is a female from HongKong. The nodes are the IPA symbols of the 9 monophthongs in Mandarin. Then, by this figure, we can see their phonetic trees are structurally very similar but slightly different. Meanwhile, the result also shows high independence of genders.

4.3 Distances between syllable-based structures

In Fig. 2, the structures are obtained by the utterances of monophthongs, which is much more steady comparing to other phonological units like syllables in Chinese dialects. In other words, more dialectal features can be shown 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 diagonal Gaussian, just like in building the monoph-

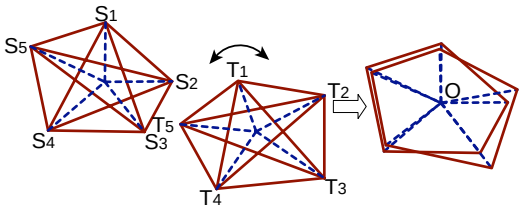

Fig. 3 Distance calculation after shift and rotation

thong structures, each monophthong segment is modeled as diagonal 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 here.

By calculating the BD of every pair of syllables, a 38×38 distance matrix can be obtained using the recorded utterances of each speaker, which fixes the unique pronunciation structure for that speaker. Then, the distance between two structures is obtained after one is shifted ($+b$) and rotated ($\times A$) until the best overlap is observed between them, which is shown in Fig. 3. 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 points of the two structures.

In [6], it was experimentally proved that the minimum sum can be approximately calculated as Euclidean distance between two distance matrices. Following is the detailed computing formula:

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

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

4.4 Experimental results and discussions

Using inter-speaker distances between dialectal syllable structures, the dialect speakers are classified using Ward's clustering method and the result is shown in Fig. 4, where every speaker is represented by speaker ID in Table 2. The dialect regions are shown by colors used in Table 2. By this figure, not only the speakers from the same dialect region are clustered together, but also the speakers from the same sub-dialect region are also clustered nearer to each other. For example, speakers 11-14 are from ShangHai and speakers 15-18 are from different cities of Wu dialect region. Meanwhile, we can see that the speakers from Min, Yue and Kejia dialects regions are clustered in a big group and this can be explained by the following facts. These three dialect regions are very close to each other geographically and it is claimed by some historical linguists that these dialect regions are affected by each other greatly during their development [12]. 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 correspond to the same speaker with different dialects. In conclusion, this result shows that these speakers are classified only by the purely linguistic information of their utterances.

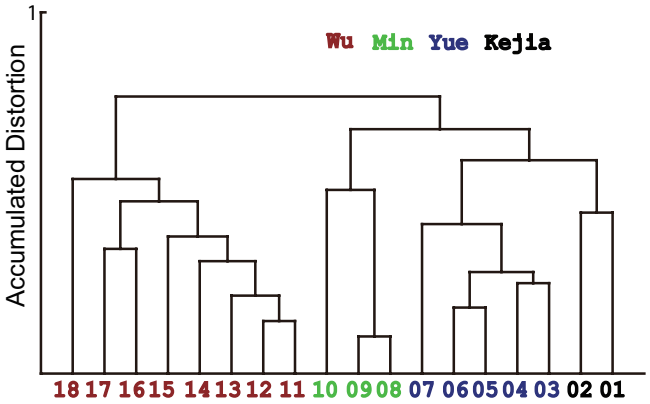

Fig. 4 Classification of the dialect speakers

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. Since different vocal tract length of speakers rotates the utterances in a cepstral space [13], 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 ($=\{a_{ij}\}$) [15].

$$a_{ij}=\frac{1}{(j-1)!}\sum_{m=m_0}^{j}\binom{j}{m}\frac{(m+i-1)!}{(m+i-j)!}(-1)^{(m+i-j)}\alpha^{(2m+i-j)},$$

where $|\alpha|\leqq 1.0$, $m_0=\max(0,j-i)$, and

$$\binom{j}{m}=\begin{cases}{}_jC_m & (j\geqq m)\\ 0 & (j<m).\end{cases}$$

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. After extracting the pitch information, the recorded data were converted into a shorter version with $\alpha=0.2$ and a taller version with $\alpha=-0.2$ using STRAIGHT [16]. Fig. 5 shows two spectrums of the same syllable produced by pseudo tall and short speakers.

5.2 Experimental results and discussions

Using the original and simulated data together, we did the same classification experiment and the result is Fig. 6. Further, another classification experiment was carried out. 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

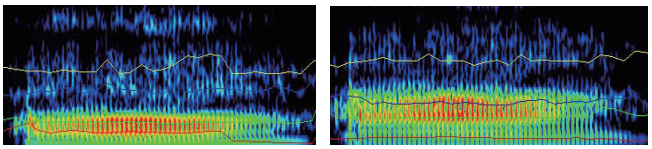

Fig. 5 Spectrums of tall and short speakers

as

$$D_2(A,B) = \sqrt{\frac{1}{M}\sum{}_i BD(S_i^A, S_i^B)}, \qquad (3)$$

S_i^A is syllable i of speaker A. Using this method with the same data set, the classification result is obtained and shown in Fig. 7. In these two figures, the original speakers are represented by the speaker IDs as Table 2, the same ID with a line on the top represents the tall version of this speaker and the same ID with a line on the bottom represents the short version. Then, in Fig. 6, 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. 4, 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 in a big group. On the other hand, in Fig. 7, which is obtained using the same data set but different distance measure of D_2, the speakers are generally classified into three big groups corresponding to their vocal tract length. These results proved that our proposed method does work well on extracting the purely dialectal and really speaker-invariant features.

6. Conclusions

In order to classify dialect speakers, we propose the use of the structural representation of pronunciation. At the beginning, a list of written characters are selected considering the phonological features of Chinese dialects, and the pronunciation structure for every speaker is built using their 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, the method is also tested and compared to conventional methods using the utterances of simulated tall and short speakers. All the results show this method can do very good and linguistically-reasonable classifications and is highly independent of extra-linguistic variations caused by speaker variability.

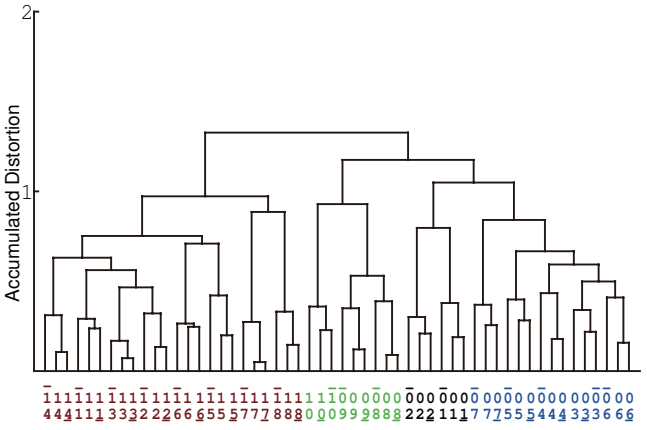

Fig. 6 Classification of adults and children using D_1

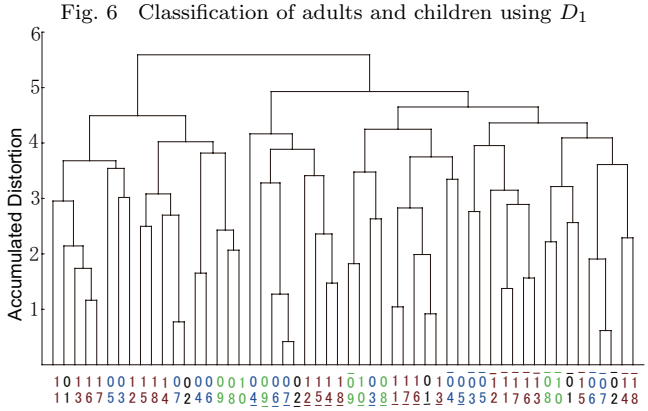

Fig. 7 Classification of adults and children using D_2

References

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

[2] 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.

[3] 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.

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

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

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

[7] 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."

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

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

[15] 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.

[16] 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.