SIMPLIFIED DEFORMATION COMPENSATION FOR EMOTIONAL SPEAKER RECOGNITION

Yingchun Yang, Tian Wu, Hongbing Lv
College of Computer Science and Technology, Zhejiang University
yycc@zju.edu.cn, wutian@zju.edu.cn, lvhbc@zju.edu.cn

ABSTRACT

Emotional speaker recognition has been investigated by a number of researchers [1-6], however, all the current approaches have flaws in the requirement of a large amount of emotional speech from speakers during training and even the emotional state of a user during testing, which hinder the commercialization of speaker recognition technology. We propose our method from novel view of MFCC deformation caused by pitch deviation, named Pitch Deviation-based Cepstral Compensation (PDC), which takes into account the correlation between glottis and vocal tract. Our method is applied to two emotional speech corpus EPS and MASC with absolute IR (Identification Rate) increase by 10.1% for the former and 4.12% for the latter, which are promising results.

Index Terms—Emotional speaker recognition, pitch deviation, cepstrum compensation, glottis-vocal tract correlation

1. INTRODUCTION

There are various challenges that can adversely impact current speaker recognition performance, including channel mismatch, background noises and intra-speaker variability, etc. Nowadays, emotional speaker recognition has attracted more and more attentions, i.e. the mismatch of the speaker’s emotion states between training and testing causes unacceptably high error rates in automatic speaker recognition (ASR) systems. Different emotion states will affect speech production mechanism in different ways, which leads to acoustical changes in his/her speech [1], thus will influence intra-speaker vocal variability and aggravate the recognition performance.

Up till now, there are mainly two ideas to cope with emotional variability. One is “structured training” proposed by University of Geneva [1], the other is treating as channel mismatch problem presented by Tsinghua University [2,3]. In the formal, the main job is to obtain emotional utterances using computer-aided induction as well as neutral ones during speakers’ training stage, thus making the system get familiar with the emotion variation of the user’s voice. Following this idea, some work have been done by Zhejiang University, including following 3 methods: 1) emotion-added model [4] that classifies the emotional states by acoustical features and generates emotion-added model based on the emotion grouping. 2) Emotion-state conversion [5], which generates the emotion models by training the emotion speech converted from neutral speech. 3) Natural-Emotion GMM which converts the neutral GMM to produce various emotional GMMs.

However, researchers of Tsinghua University have proposed an interesting solution, treating the emotional variability as the familiar channel mismatch problem. They have given 2 methods: 1) An emotion-dependent score normalization method [2] derived from H-norm is proposed to alleviate the effect of speaker emotion variability. 2) Emotion Attribute Projection [3] derived from NAP, removes from the SVM expansion dimensions that are irrelevant to the speaker recognition problem on emotional speech.

Compared with traditional method training only with neutral utterances, above approaches can enhance the speaker recognition performance to some degree. However, all these approaches required a large amount of emotional speech from speakers during training and even their emotional states during testing, which remain as a great difficulty in real-life situation, thus construct a barrier for speaker recognition commercialization.

In this paper, we try to regard emotional variability from the novel view of deformation, i.e. spectral features distribution deformation caused by pitch deviation. Therefore, our idea is to neutralize the speech features by inverse compensation according to pitch variability.

The remainder of this paper is organized as follows: the next section describes the motivation of this research. It is followed by the proposed model including spectral feature compensation and GMM solution in section 4. The experiments are given in section 5. Finally, Section 6 presents the conclusions.

2. MOTIVATION

2.1. Related work

As is well known, spectral parameters, such as Mel Frequency Cepstral Coefficient (MFCC), Linear Prediction Coefficient (LPC), encode vocal tract information. While pitch (the vibration frequency of the vocal fold, fundamental frequency) is known to be an important feature to characterize articulation source (glottis). Assuming the short-term independence of the glottis and the vocal tract, most speaker recognition systems employ MFCC and usually ignore the contributions of the glottis [7,8]. Nobuki firstly [9] analyzed the correlation between spectral variations and F0 changes in a vowel sound, and applied the correlation to improve speech recognition. [7] study the correlation between the glottis and vocal tract by generating speaker sub-models for each pitch
range, and observed an increase in the speaker identification rate. These work showed that the dependence of glottis and vocal tract do exist and can be useful for speech/speaker recognition. However, the mechanisms involved in speech production are complex and imply complex dependence of articulators and vocal folds, which make the correlation modeling work tough.

Emotion detection and classification is a hot topic. Some research work mentioned mean fundamental frequency as a good indicator for emotion classification[10], which give us the idea of utilizing mean fundamental frequency as indicator to tell emotion from neutral.

### 2.2. Our observation

To illustrate the association of variation of MFCC and fundamental frequency under different emotion. We chose 4 emotional utterances from a male speaker, Chris Cantellini, obtained from EPS database that is described in section 5. Each utterance last about 30 seconds and is used to extract fundamental frequency vectors and the corresponding MFCC vectors. Here pitch is determined by SHR (Subharmonic-To-Harmonic Ratio) method provided by Sun[11]. The pitch range is divided into 20 linearly bins of 25Hz width.

Figure 1 exhibits the pitch histograms of four kinds of emotion, neutral, anxiety, elation and hot anger. It can be seen that the histograms are different and vary from one kind of emotion to another. Figure 2 shows the variation of corresponding MFCC distribution under 4 emotion states. As is well known, Mel Frequency Cepstral Coefficients are Gaussian distributed. It is showed on the figure that for emotional states, the mean of the second and the third dimension of MFCCs have an excursion compared with the neutral. On the other hand, variations of the distribution's shape also exist. With the help of the figures, it is observed that the pitch and MFCC distributions do vary under different emotions. The main variation includes the deviation of pitch mean and MFCC mean. Similar variations are also detected in other order of MFCC as well as those of other emotional speakers. Consequently, emotional states changes do have influence on both MFCC and fundamental frequency, which can increase intra-speaker variability that is adverse to speaker recognition task.

**Fig. 1. Pitch histograms under different emotional states.**

**Fig. 2. The 2nd and 3rd MFCC distribution for the different emotional states of one speaker.**

### 3. PROPOSED MODEL

In terms of the phenomenon of pitch and MFCC’s distribution variation that is presented in last section, a Pitch Deviation-based Cepstrum Compensation(PDCC) model is proposed. It is supposed that changes of fundamental frequency around its expectation will lead to deformation of cepstrum features, without which MFCCs could represent neutral speaker more exactly.

In section 2, it is observed that speaker features distributes diversely in different emotional states, which will influence the speaker recognition system. As is shown in Figure 2, one of the major deformations of MFCCs in different emotional states is mean excursion. The initial research on spectral feature compensation can be focused on the linear model. For the speaker recognition on affective speech, two assumptions are introduced here:

A. Deformations of cepstrum features are related with fundamental frequency while emotional states are changing.

B. One of the most important deformation of MFCCs in various emotional states is mean excursion, without which MFCCs could represent neutral speaker more exactly.

For a certain segment of a speech, MFCCs and pitch values are denoted by $x(t)$ and $y(t)$ respectively. In the preprocessing, frame length is set to $\Delta T$. $\tilde{x}(t)$ represents estimated MFCC vector estimated from a windowed signal centered at time $t\Delta T$, and $y(t)$ is the estimated pitch also estimated at time $t\Delta T$. For each estimated $\tilde{x}(t)$, we assume that it corresponds to an deformation compensation representation $x_d(t)$, which is considered to better characterize the neutral speaker. The deformation compensation MFCC feature can be assumed as:
\[ \ddot{x}(t) = \ddot{x}(t) - \ddot{\alpha} \times \frac{|y(t) - E(y(t))|}{|E(y(t))|} \]  

(1)

Here, the parameter \( \ddot{\alpha} \) is the deformation factor which has the same dimension as MFCC vector \( \ddot{x}(t) \). The deformation factor describes the assumed linear correlation between pitch and MFCC, and we regard it as an important speaker characteristic. \( E(\cdot) \) is the expectation function.

4. EXPERIMENTS

4.1. Speech database

The two emotional corpus we utilized can be obtained from Linguistic Data Consortium (LDC). One corpus we used is Emotional Prosody Speech (EPS) corpus provided by LDC. This corpus is recorded by 8 English speakers, including 3 men and 5 women who are actors and actresses. The categories of emotions are chosen based on Banse & Scherer’s selection criteria [12]. They are drawn from over 3,800 descriptions of emotional experiences as well as the situations from which these responses are elicited [1,12]. Each speaker read 10 to 40 sentences of different types portraying the emotional states. For each person, the database contains 5-minute emotional speech and 0.5-to-1-minute neutral speech.

The other corpus is MASC (Mandarin Affective Speech Corpus) collected in our lab [12]. 68 Chinese students from Zhejiang University were selected to express five emotional states. Each speaker reads 5 phrases, 20 sentences three times for each emotional state and 2 paragraphs only for neutral. The materials of this corpus cover all the phoneme in Mandarin [13].

The speech signal is segmented into 32 ms with 16 ms overlapped frames, pre-emphasized and Hamming windowed. 32 MFCC (Mel Frequency Cepstrum Coefficients) [13] is used to describe the speakers’ characteristics, while 32 GMMs (Gaussian Mixture Model) is employed as the classifier. SHR (Subharmonic-To-Harmonic Ratio) algorithm [11] is used for extracting pitch. IR (Identification Rate) and EER (Equal Error Rate) are chosen as the criteria of performance.

4.2. The Baseline Strategy

The baseline strategy uses traditional MFCC-GMM in the experiment. Only MFCC without compensation are utilized. GMM are trained with EM algorithm on neutral speech and tested by emotional speech. This experiment is used for comparison.

4.3. PDCC Strategy

For the second strategy, Pitch Deviation-based Cepstrum Compensation (PDCC) are used. For training stage, besides constructing traditional GMM speaker model same to the baseline strategy, the deformation factor \( \ddot{\alpha} \) in equation (1) need to be estimated. We suppose each dimension of the factor is independent of others. In order to get each dimension of \( \ddot{\alpha} \), local optimal solution is used. The properly compensated MFCC features could make the system's performance maximal, i.e. obtaining maximum posterior probability. Neutral speech of each speaker is divided into two parts, one for training and the other for evaluation to estimate \( \ddot{\alpha} \). For testing stage, each MFCC vector is compensated according to equation (1) before inputting to each speaker model. Here, compensation zone \( t \) is restricted to each natural utterance (such as twenty two) for EPS corpus, and 30 second segment in neutral paragraphs for MASC corpus.

4.4. Experimental results

In order to find out whether our method is effective in broad domain. We apply our method to two emotional corpus from different culture and with different style. EPS is from professional English actors and actresses, and they mainly pronounced digit utterances, eg. “July Eighteenth”, “Two thousand nine”. While MASC is from ordinary Chinese students, and they mainly pronounced Mandarin phoneme balanced sentences, eg. “Ni Shi Ge Hao Ren(You are a good man)”, “Jin Tian Wan Shang Hui Xia Yu(It will rain this evening)”. There are important points deserve our considerations. Firstly, our proposed model obtain lower EER and higher IR than baseline in all experiments. IR increase by 10.1% for EPS corpus, and increase by 4.12% for MASC corpus (ref. to Tab. 1 and 2). It is obvious that the deformation compensated features behave better than the unsettled ones.

<table>
<thead>
<tr>
<th>Method</th>
<th>EPS(%)</th>
<th>MASC(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>62.94</td>
<td>37.78</td>
</tr>
<tr>
<td>PDCC</td>
<td>73.04</td>
<td>41.90</td>
</tr>
</tbody>
</table>

Secondly, our proposed model obtain different degree of improvement in IR for different emotion. Figure 3 indicates the IRs in each emotional state on EPS. The increases of 14 emotional states are quite different. The highest performance raise is on panic, which shows 19.86% increase in IR. Figure 4 indicates the IRs in each emotional state on MASC. The increases of 4 emotional states are quite different. The higher performance raise are on panic and elation. This phenomenon give us hint that we should adopt different deformation compensation for different emotion.
Thirdly, our proposed model obtain different degree of improvement in IR for different corpus. It is observed from the Figure 3 that the PDCC line of IR is far above the baseline, corresponding to IR increase by 10.1%. However, the PDCC line of IR is slightly above the baseline, corresponding to IR increase by 4.12%. It’s supposed that the reasons are two: 1) Speakers of EPS are professional actors and actresses and can pronounce genuine emotional speech, which is more suitable for PDCC than MASC. 2) MASC adopted Mandarin, a tone language, which has pitch variability with different contents itself, thus make pitch deviation confuse by both contents and emotion change.

![Fig. 3. IRs of each emotional states on EPS.](image)

![Fig. 4. IRs of each emotional states on MASC.](image)

6. CONCLUSIONS

In this paper, we propose a novel approach by MFCC compensation according to pitch deviation for emotional speaker recognition. Our utilizing the correlations of MFCC and fundamental frequency (pitch) is quite different from other current approaches. Experiments are conducted on two emotional speech database and the results are encouraging.

As future work, besides comparing our method with other emotional speaker recognition methods, looking for more proper and accurate model for compensating MFCC by fundamental frequency should be the first problem. Why the current speaker recognition system is confronted with different degree of performance degradation for different emotion states? How to adopt different deformation compensation for different emotional states is the second problem. Finally, our current method only consider the voiced speech, how to deal with the unvoiced speech is the third problem.

7. ACKNOWLEDGEMENTS

This work is supported by the following foundings: 863 Program 2006AA012Z136, ZJNSF_Y106705 and NSTPP 2006BAH02A01

8. REFERENCES