Emotional Development Method of Virtual High Differences Emotional Speech Based on the Baseband Mapping

Hongbo Zhang, Teng Wang, Xiang Yang*

School of Physics & Electrical Information Engineering, Ningxia University, Yinchuan, 750021, China
*Corresponding author, Ningxia University, tel-13469582532, email-2720695@qq.com

***Abstract*—Emotion mismatch between training and testing is one of the important factors causing the performance degradation of speaker recognition system. In this paper, we proposed an emotional development method based on virtual high differences emotional speech of the baseband mapping to build the speech closer to the real high differences emotional speech on the characteristics distribution, by adjusting the neutral voice baseband mean, then build the speakers' high difference emotional model. And by combining with the neutral model to improve speech description ability of speaker model in every emotional state. As a result, the recognition performance of emotional speaker will be improved.**

Keywords-baseband mapping; score reliability fusion; emotional speech; emotional expanding

I. INTRODUCTION

In the most of study about speaker recognition technology, the changes of environment or channel which is something about robustness is considered most. Less research work to consider the effect of speaker's own change such as their mood. Emotion mismatch between training and testing will cause system performance decline sharply which is emotional speaker recognition. K. R. Scherer [1] presents a structured training method, by collecting many registered users' speech in different emotional state, build a model which cover the whole space of speaker's voice characteristics distribution to solve this problem. But in practical applications, Systems generally could only get the user's neutral training voice. Therefore it is difficult to implement the structured training method. In order to avoid speaker emotional speech additional requirements, Li Dongdong [2] use the differences in prosodic features of speech which is from part of people pronunciation for the same content in the different emotional states as a neutral and emotional speech characteristics conversion law, take neutral speech features translate into speech features in different emotional states to train the speaker model in different emotional states. Shan Zhenyu [3] proposed an emotion model conversion method which use mapping rule of emotional speaker model and neutral speaker model parameter to build testers' emotional model. These two methods improve the system performance in different extent, But both need to select the speaker's corresponding emotion model to match and calculate score according to the test voice emotion category, that need the support of speech emotion recognition. However, the effect in speaker-independent speech emotion recognition is not ideal, such as the emotion recognition rate of multiple speaker in five emotions by Ververidis [4] was 53%.

In this paper, we use polynomial fitting method to fit The baseband mean value variation rule of neutral speech and high differences emotional speech, then predict the baseband mean value of speaker's high differences emotional speech in test set, and by adjusting the baseband mean of neutral voice to change indirectly its channel characteristics to build a virtual high differences emotional speech.

II. THE STRUCTURE OF VIRTUAL HIGH DIFFERENCES EMOTIONAL SPEECH BASED ON THE BASEBAND MAPPING

Using function fitting method to study The baseband mean value variation rule of neutral speech and high differences emotional speech in development data, and map the baseband mean value of speaker's high differences emotional speech. Last we can build a virtual high difference emotional training voice by adjusting the neutral voice baseband mean for each speaker.

A. The structure of high differences emotional speech baseband sequence.

Speakers usually express their emotions(such as angry, happy and Scared) by changing the tone (baseband mean). When people express the particular emotion, baseband mean relative to the Changes amplitude of neutral speech is generally related to their vocal cords characteristics (Commonly used baseband mean to describe). There is a big difference on the emotion expression between male and female. Here we delimit that the baseband changes amplitude is $f_g(\overline{L})$ when the speakers express the high differences emotion. $\overline{L}$ is the baseband mean of the neutral speech, g is the gender information. If we know the f_g function definition, we can use the formula(1) to adjust the baseband mean of the neutral speech, thus we can get the baseband sequence of the high differences emotional speech.

$$H_t = f_g(\overline{L}) * L_t \tag{1}$$

H_t was the pitch of a period of emotional speech, while L_t was the normalized pitch that approximated the speaker's corresponding neutral speech.

Obviously, the form of f_g was unknown and hard to solve analytically. Polynomial was a smooth and continuous function, and its differential form was also a polynomial. So

it was a good choice for polynomial to fit f_g. Polynomial form was as follow:

$$f_g(x) = \sum_{i=0}^{p_g} a_{ig} x^i \tag{2}$$

Here, a is the coefficients of polynomial, p is the order of polynomial, g is gender.

Here, we use AIC criterion [5] to determine the order of polynomial function, which make the AIC reach to the minimum. In this paper, we use the simplified form of AIC criterion:

$$AIC = 2m + n[\ln(RSS / n)] \tag{3}$$

The m is the parameters number ($m = p + 1$)of fitting function, the n is the number of observed sample, and RSS is the residual sum of squares $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$.

B. The determining of polynomial coefficients

If the polynomial order p has been set, we could use least squares to calculate the specific parameters of polynomial fitting function in this order. For the given training data $(x_i, y_i)(i = 0,1,\mathrm{L},n)$ (here, x_i is the baseband mean of the speakers' neutral speech. y_i is the high differences emotional baseband mean of the same speaker). It demand to find a function $y = f^*(x)$ in the kind of function $\varphi = \{\varphi_0, \varphi_1, \mathrm{L}, \varphi_p\}$ ($\varphi_j = x^j$), which to make the error sum of squares (formula 4) be minimum.

$$\|\delta\|_2^2 = \sum_{i=0}^{n} \delta_i^2 = \sum_{i=0}^{n} [f^*(x_i) - y_i]^2 = \min_{f(x) \in A} \sum_{i=1}^{n} [f(x_i) - y_i]^2 \tag{4}$$

Here, $f(x) = a_0 + a_1 x + \cdots + a_p x^p \ (p < n)$

It can transformed into solving the minimum point $(a_0^*, a_1^*, \mathrm{L}, a_p^*)$ of multivariate function:

$$I(a_0, a_1, \cdots, a_p) = \sum_{i=0}^{n} [\sum_{j=0}^{p} a_j \varphi_j(x_i) - y_i]^2 \tag{5}$$

According to the requirement of multivariate function extremum, we can get

$$\frac{\partial I}{\partial a_k} = 2 \sum_{i=0}^{n} [\sum_{j=0}^{p} a_j \varphi_j(x_i) - y_i] \varphi_k(x_i) = 0 \ (k = 0,1,\cdots,p) \tag{6}$$

If we mark $(\varphi_j, \varphi_i) = \sum_{i=0}^{n} \varphi_j(x_i)\varphi_i(x_i)$, $(y_i, \varphi_k) = \sum_{i=0}^{n} y_i \varphi_k(x_i) \equiv d_k$ $(k = 0,1,\mathrm{L},p)$, we can get

$$\sum_{j=0}^{n} (\varphi_j, \varphi_k) a_j = d_k \ \ (k = 0,1,\cdots,p) \tag{7}$$

This function is called normal equation, its matrix form is

$$Ga = d \tag{8}$$

Here, $a = (a_0, a_1, \mathrm{L}, a_p)^T$, $d = (d_0, d_1, \mathrm{L}, d_p)^T$,

$$G = \begin{bmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) & \cdots & (\varphi_0, \varphi_p) \\ (\varphi_1, \varphi_0) & (\varphi_1, \varphi_1) & \cdots & (\varphi_1, \varphi_p) \\ \cdots & \cdots & \cdots & \cdots \\ (\varphi_p, \varphi_0) & (\varphi_p, \varphi_1) & \cdots & (\varphi_p, \varphi_p) \end{bmatrix} \tag{9}$$

Because of $\varphi_0, \varphi_1, \mathrm{L}, \varphi_p$ linearly independent, $|G| \neq 0$, solutions of equations is unique.

$$a_k = a_k^* \ (k = 0,1,\cdots,p) \tag{10}$$

Then we can get the least squares solution of the function $f(x)$.

$$f^*(x) = a_0^* \varphi_0(x) + a_1^* \varphi_1(x) + \cdots + a_p^* \varphi_p(x) \tag{11}$$

It is provable that $\sum_{i=0}^{n} [f^*(x_i) - y_i]^2 \le \sum_{i=0}^{n} [f(x_i) - y_i]^2$

So $f^*(x)$ is the least squares solution.

C. Synthesis of virtual heigh difference emotional speech

We belive that there is mutual interference phenomena between sound source and channel. [6,7]So to some extent we can speculate that the change of the sound source feature （just like f0） will change the channel(just like MFCC). This paper, we get the virtual high differences emotional speech by adjusting the baseband distributed center of neutral speech tend to the fitting high differences baseband mean. Figure1 describe the synthetic process of virtual high differences emotional speech. This process divided into three steps: 1) Determine the order of male and female polynomial fitting function according to AIC standard; 2) Learning the male and female baseband mapping function with the speech data in training set; 3) For given neutral speech, extracting the baseband sequence by using autocorrelation algorithm [8] firstly, then obtaining baseband sequence of the corresponding high differences emotional speech according to formula 1. At the last, re-synthesis the modified baseband sequence and the original speech into new virtual high difference emotional speech with PSOLA [9] method.

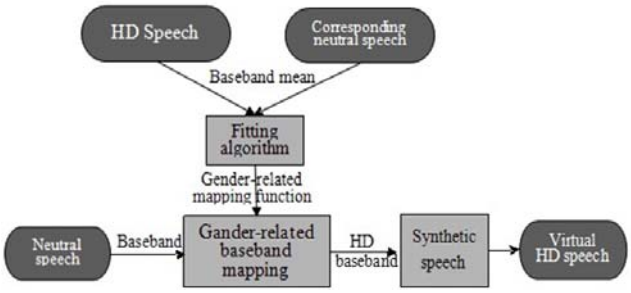

Figure 1 Synthetic framework of virtual high differences emotional speech

III. Fusion Weight Estimating Strategy based on Recognition Rate

Synthetic virtual high differences emotional speech is different from the real emotional speech, so there is unreliability in the score which is got from the virtual high differences model λ_H. While the score which is got from X_L on the neutral model λ_N is reliable. For the two kinds different reliability score, it is unreasonable obviously to plus equal weight.

Determine the model collection $Z_\theta = \{\lambda_{\theta i} | i = 1,2,\mathrm{L},M\}$. $\theta \in \{H,N\}$, M is the number of registration speaker, H is the type of high differences, N is the neuter. Testing speech collection is $\mathrm{O}_\varphi = \{X_{\varphi j} \mid j = 1,2,\cdots,K_\varphi\}$, $\varphi \in \{H,L\}$, K_φ is the φ kind number of testing speech, L is the type of low differences.

When the speakers use collection Z_θ to determine the speech identity in the O_φ collection in the recognition, the higher the speakers identification rate $IR_{\theta\varphi}$, the higher the proportion which the testing speech of O_φ is identified correctly by model collection Z_θ Similarly, the match score of speech in the O_φ on the model which is in Z_θ is more reliable. According this, we can determine the weight α and β of formula (12):

$$\begin{cases} \alpha = \dfrac{IR_{NL}}{IR_{NL} + IR_{HH}} \\ \beta = \dfrac{IR_{HH}}{IR_{NL} + IR_{HH}} \end{cases} \tag{12}$$

IR_{NL} and IR_{HH} in formula (12) can express as:

$$IR_{NL} = \frac{NUM_{L\_right}}{NUM_{L\_Total}} \tag{13}$$

$$IR_{HH} = \frac{NUM_{H\_right}}{NUM_{H\_Total}} \tag{14}$$

In the formula, NUM_{L_right} is the number that speech of O_L is distinguished correctly by collection Z_N, NUM_{L_Total} is the total number of speech in O_L. And NUM_{H_right} is the number which the number that speech of O_H is distinguished correctly by collection Z_H, NUM_{H_Total} is the total of speech in O_H.

IV. Experimental Analysis and Discussion

A. Experimental setting

The corpuses used in the experiment were split into 3 parts: development data (Speeches of the first 18 people in MASC), test data (Speeches of the last 50 people in MASC) and testing speeches (speeches of 7 speakers in EPST corresponding with 5 same emotional classifications as MASC). In the experiment, The weight coefficients α and β, baseband mapping function f, and gender models are all got from the dates in development data. The traditional GMM-UBM was setup as compared baseline. UBM was adopted 1024 order and characteristics were 13-dimensional MFCC and its delta. The length of window for MFCC, energy and pitch were 32ms uniformly, and step sizes were 16ms uniformly.

B. The optimal order of polynomial fitting function

When we use polynomial fitting baseband mapping function f_g, we need to determine polynomial order p first, then use the least squares to determine the specific forms.

In order to determine order p, we make $p = 1,2,\mathrm{L},20$, and fit the corresponding polynomial function. And then select the p which make the AIC value minimum as optimal order. In this experiment, training corpora use all the speech of the seven male and seven female in the development data. The relation curve between polynomial fitting function order and AIC was depicted in Figure 2, from which we can know order p is 11 for male, and 5 for female.

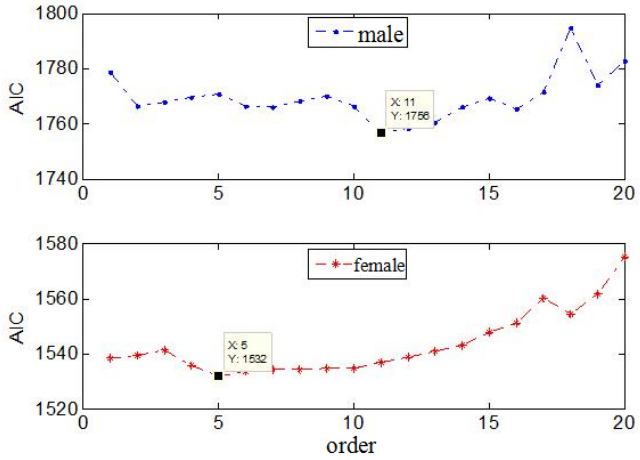

Figure2 The relation between the order of polynomial function and the AIC

C. Experimental analysis and discussion

For clearly expressing MFCC feature distributed change under the emotional change and virtual synthesis HD emotional speech effect. We choose one speaker randomly, and analysis the speaker's neutral speech ($f0 = 210.73\mathrm{Hz}$) and distributed differences in MFCC between HD emotional speech (angry emotion $f0 = 291.33\mathrm{Hz}$) and virtual synthesis high differences emotional speech ($f0 = 288.59\mathrm{Hz}$). Figure 3 express the distributed situation of MFCC feature of these three speeches in 8, 9, 13 three-dimensional.

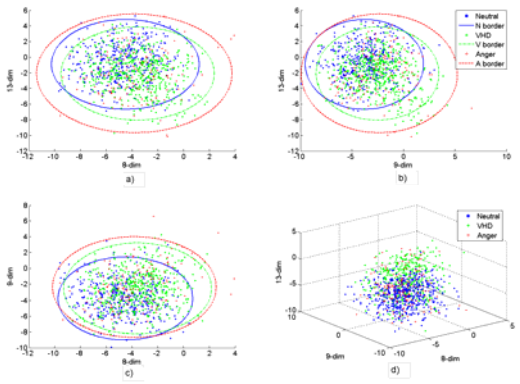

Figure 3 The effect of virtual high difference synthetic speech based on time-frequency mapping

From the figure, we can see that there is an apparent distance between every dimensional distributed center of neutral speech feature and anger speech. In addition, anger speech feature is more discrete than neutral speech. The speech distributed center by virtual synthesis and dispersion are closer to real HD emotional speech (Anger). In summary, synthesis virtual HD emotional speech feature is really closer to the real HD emotional speech than neutral speech. And virtual construction HD emotional speaker model is more effective to describe HD emotional speech feature distribution than neutral speaker model.

V. Conclusion

In this paper we build a virtual high differences emotional Speech model based on the baseband mapping, which is combined with neutral model, to improve the ability that speaker model express the multiple emotional statute speech. And improve the identification performance of emotional speakers. When we calculated the score of test voice high mismatch part in its high-difference model and low mismatch part in neutral model respectively, we use weights evaluation algorithm to determine the weights of two group scores. Compare to the experiment result of the other weights strategy, for bi-model method, the weights coefficient by weights evaluation algorithm achieve a better performance than that by using the equal weight, and the better one even achieves a result comparable to that by using the best weights selected by exhaustive strategy.

Acknowledgment

This research was supported by the Natural Science Foundation of Ningxia Hui Autonomous Region, China (Grant No. NZ1139), and Scientific and technological projects in Ningxia (The research and development application demonstration of Ningxia milk and the products' safety traceability information system which is based on the Internet of Things). All supports are gratefully acknowledged.

References

[1] Scherer K. R., Johnstone T., Klasmeyer G., et al. Can automatic speaker verification be improved by training the algorithms on emotional speech?[C]. ICSLP, 2000,2: 807-810.

[2] Li D., Yang Y., Wu Z. Emotion-State Conversion for Speaker Recognition[C]. Proceedings of ACII, Beijing, 2005: 403-410.

[3] Shan Z., Yang Y., Ye R. Natural-Emotion GMM Transformation Algorithm for Emotional [C]. INTERSPEECH, 2007: 782-785.

[4] Ververidis D., Kotropoulos C., Pitas I. Automatic emotional speech classification[C]. ICASSP, Montreal, 2004,1: 593-596.

[5] Akaike H. A new look at the statistical model identification[J]. Automatic Control, IEEE Transactions on, 1974,19(6): 716-723.

[6] Childers D. G., Yea J. J., Bocchieri E. L. Source/vocal-tract interaction in speech and singing synthesis[J]. Proc Stockholm Music Acoust Conf, 1983: 125-141.

[7] Childers D. G., Wong C. F. MEASURING AND MODELING VOCAL SOURCE-TRACT INTERACTION[J]. Biomedical Engineering, IEEE Transactions on, 1994,41(7): 663-671.

[8] Rabiner L. On the use of autocorrelation analysis for pitch detection[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1977,25(1): 24-33.

[9] Moulines E., Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones[J]. Speech Communication, 1990,9(5-6): 453 - 467.