*Journal of Signal and Information Processing*, **2013, 4, 1-6**
doi:10.4236/jsip.2013.43B001 Published Online August 2013 (http://www.scirp.org/journal/jsip)

Scientific Research

# Applying Score Reliability Fusion to Bi-Model Emotional Speaker Recognition

**H. B. Zhang[1], T. Wang[1], T. Huang[2], X. Yang[1]**

[1]School of Physics & Electrical Information Engineering, Ningxia University, Yinchuan, 750021, China; [2]MicroStrategy Software (Hangzhou) Co., Ltd., Hangzhou, Zhejiang, 310012, China
Email: zhb@nxu.edu.cn

Received March, 2013.

## ABSTRACT

Emotion mismatch between training and testing is one of the important factors causing the performance degradation of speaker recognition system. In our previous work, a bi-model emotion speaker recognition (BESR) method based on virtual HD (High Different from neutral, with large pitch offset) speech synthesizing was proposed to deal with this problem. It enhanced the system performance under mismatch emotion states in MASC, while still suffering the system risk introduced by fusing the scores from the unreliable VHD model and the neutral model with equal weight. In this paper, we propose a new BESR method based on score reliability fusion. Two strategies, by utilizing identification rate and scores average relative loss difference, are presented to estimate the weights for the two group scores. The results on both MASC and EPST shows that by using the weights generated by the two strategies, the BESR method 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.

**Keywords:** Emotional Speaker Recogitnion; Score Reliability Fusion; Fusion Weight Estimating Strategy; Bi-Model

## 1. Introduction

In the most studies about speaker recognition technology, the changes of environment or channel which are 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.

In order to avoid this problem in speech emotion recognition, in this paper we propose a weight strategy based on scores average relative loss difference. Using scores average relative loss difference of various types of testing voice on the corresponding model set to estimate the various types of score weight coefficient. In addition, the weight strategy based on the recognition rate using the two models on the respective classification test speech recognition rate as a test class model score weighted right respectively. The results on both MASC and EPST [1] show that using the Bi-model system of fusion weight estimating strategy achieve a better performance than that by using the equal weight.

## 2. Applying Score Reliability Fusion to Bi-Model Emotional Speaker Recognition

In this method, we mark the emotional speech which is different from the neutral voice on MFCC and baseband such as angry, happy and scared state as high different voice. And we build a virtual high difference in emotional training voice by adjusting the neutral voice baseband mean for each speaker, Thereby reducing the degree of mismatch between the test voice and training model. **Figure 1** describes the system framework of method. During the training process, we establish the two models for each speaker: Using the synthetic virtual high difference emotional speech to train high difference model, the neutral speech for neutral model. During the testing process, we can get the gender information of test voice from gender recognition. Then mark the high mismatch part by using gender-dependent mismatch detection. Finally, for each speaker, calculated the score of test voice high mismatch part in its high-difference model and other low mismatch part in neutral model respectively, fuse them using linear weighted fusion.

### 2.1. Gender Recognition

Gender recognition can be regarded as a special speaker recognition that speaker is divided into two types: male and female. Male gender model is trained by male speaker's corpus $M_m$. Female gender model is trained by female speaker's corpus $M_f$. When testing, match scores

 

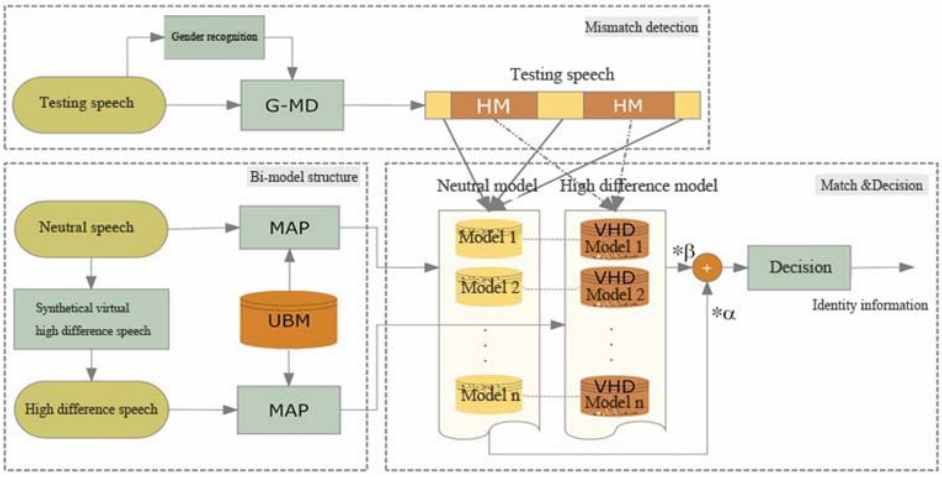


**Figure 1. Applying score reliability fusion to bi-model emotional speaker recognition.**

between testing speech X and the 2 gender models were computed, and gender of the model with the highest score was the speaker's gender.

## 2.2. High Mismatch Detection

Our previous study [2] found three high differences emotional speech baseband mean relative mean neutral speech Base frequency (the same speaker and the same text) there is a certain deviation, the greater the deviation with the lower the degree of matching of the speech, the lower the probability of correctly recognized. Accordingly, we mark higher baseband mean of high differences emotional speech as high mismatch. Because of the big differences between male and female speech, the male and female are matched by training in the mismatch detection. Specific operations can be divided into two steps: Firstly, we use the differences detection technology to test weather the Statement belongs to high differences. In the differences detection technology, we use the classification method based on KNN and we have chose eight global features of the speech snippet: baseband and the mean, the maximum, the minimum and standard deviation of log domain energy. Secondly, we divide the high different emotional speech into several snippets. The next, the baseband mean which is higher than the threshold is marked as the high mismatch parts. The threshold is the value of error rate (EER) point, by using baseband mean to distinguish between low differences(neutral and sad) and high differences(angry, happy and scared) speech in parameter development set (male:156 Hz, female: 250 Hz).

## 2.3. The Virtual High Differences in Emotional Speech Construction Based on Time-Frequency Mapping

Between the sound source and channel interference phenomena [3,4], the change of the sound source characteristics (f0) It can be speculated to some extent led channel characteristics (such as MFCC) change. It tends to baseband mean of high differences emotional speech by adjusting baseband distribution center of the neutral speech, resulting in the virtual high differences in emotional speech. When the 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.

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

$L_t$ is the baseband value of neutral speech and $H_t$ is the baseband value of the virtual high differences emotional speech at the t frame.

But the form of $f_g$ is unknown , and it's difficult to get an analytic solution. Here we use polynomial function to fit $f_g$ . At the same time, 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:

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

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

When the $f_g$ is known, we can use the autocorrela-

 

tion algorithm [6] to extract the baseband from the neutral speech, then get the baseband sequence of "the high differences emotional speech" according to formula(1). At last, we can get the corresponding virtual high differences emotional speech through correcting the baseband by PSOLA method [7].

### 2.4. Emotional Speaker Recognition

This system is built on the frame foundation of GMM-UBM. Every registered user i adaptive to two son model from UBM ( $\lambda_{ubm}$ ) . Neutral speech adaptive to neutral model ( $\lambda_{iN}$ ). Synthetic virtual high differences emotional speech adaptive to $\lambda_{iN}$ . The testing speech $X=\{x_t \mid n=0,1,\cdots,T-1\}$ is divided high mismatch part ( $X_H, H=\{t \mid t=j_0, j_1,\cdots, j_k, 0\le k<T, 0\le j_k<T\}$ ) and low mismatch part ( $X_L, L=\{t \mid 0\le t<N, n\notin H\}$ ) firstly during the test. Next, we can calculate score of the testing speech X on registered user i model by formula (3).

$$\Lambda_i(X)=\frac{1}{T}(\alpha\log\frac{p(X_L\mid\lambda_{iN})}{p(X_L\mid\lambda_{ubm})}+\beta\log\frac{p(X_H\mid\lambda_{iH})}{p(X_H\mid\lambda_{ubm})}) \tag{3}$$

$\alpha$ and $\beta$ are the weights fused score of $X_L$ on the $\lambda_{iN}$ and $X_H$ on the $\lambda_{iH}$ , $\alpha+\beta=1$ .When $\alpha=\beta=0.5$, which is so-called bi-model way of equal weight [2]. In this paper, we use two kinds of fusion weight estimating strategy based on the score reliability assessment to determine $\alpha$ and $\beta$. At last, we judge the testing speech belong to the highest score speaker in the $\Lambda(X)$.

## 3. Fusion Weight Estimating Strategy Based on the Score Reliability Assessment

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 $\lambda_H$. While the score which is got from $X_L$ on the neutral model $\lambda_N$ is reliable. For the two kinds different reliability score, it is unreasonable obviously to plus equal weight. In this part we will propose two fusion weight estimating strategy based on the score reliability assessment.

### 3.1. Based on Fusion Weight Strategy of Scores Average Relative Loss Difference

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

In the speaker recognition, we determine the testing speech belong to the speaker who correspond the model of the maximum matching probability values. The score of testing speech $X_i$ which is on the model $\lambda_i$ can be expressed as:

$$S_{ij}=\log p(X_j\mid\lambda_i) \tag{4}$$

Suppose the testing speech $X_{\varphi i}$ is the speech of speaker $\hat{i}$, the model collection of speaker recognition system is $Z_\theta$. If

$$\hat{i}=\arg\max_{1\le i\le M}(\log p(X_{\varphi j}\mid\lambda_{\theta i})) \tag{5}$$

this system can distinguish the testing speech $X_{\varphi i}$ correctly. On the contrary, the bigger the distance between the score $\log p(X_{\varphi i}\mid\lambda_{\theta\hat{i}})$ and the maximum collection score the worse the ability of the system to distinguish speech.

When we use the model collection $Z_\theta$ determine identity of the speech which is in the collection $O_\varphi$, scores average relative loss difference can be determined to :

$$C_{\theta\varphi}=\frac{1}{K_\varphi}\sum_{j=1}^{K_\varphi}\frac{\max_{\lambda_i\in Z_\theta}(\log p(X_{\varphi j}\mid\lambda_i))-\log p(X_{\varphi j}\mid\lambda_{\theta\hat{i}})}{\max_{\lambda_i\in Z_\theta}(\log p(X_{\varphi j}\mid\lambda_i))-\min_{\lambda_i\in Z_\theta}(\log p(X_{\varphi j}\mid\lambda_i))} \tag{6}$$

$\max_{\lambda_i\in Z_\theta}(\log p(X_{\phi j}\mid\lambda_i))$ and $\min_{\lambda_i\in Z_\theta}(\log p(X_{\phi j}\mid\lambda_i))$ are the maximum and minimum which the score of $X_{\varphi j}$ on the model collection $Z_\theta$. $\lambda_{\theta\hat{i}}$ is the model of $X_j$ belong to speaker in $Z_\theta$. The bigger the $C_{\theta\varphi}$, the more unreliable the score of the testing speech collection $O_\varphi$, which is on the model collecting $Z_\theta$. We can use $C_{\theta\varphi}$ to estimate $\alpha$ and $\beta$ in formula (3):

$$\begin{cases}\alpha=\dfrac{1/C_{NL}}{1/C_{HH}+1/C_{NL}}\\ \beta=\dfrac{1/C_{HH}}{1/C_{HH}+1/C_{NL}}\end{cases} \tag{7}$$

### 3.2. Based on the Weight Strategy of Recognition Rate

When the speakers use collection $Z_\theta$ to determine the speech identity in the $O_\varphi$ collection in the recognition, the higher the speakers identification rate $IR_{\theta\varphi}$, the higher the proportion which the testing speech of $O_\varphi$ is identified correctly by model collection $Z_\theta$ Similarly, the match score of speech in the $O_\varphi$ on the model which is in $Z_\theta$ is more reliable. According this, we can determine the weight $\alpha$ and $\beta$ of formula (3):

$$\begin{cases}\alpha=\dfrac{IR_{NL}}{IR_{NL}+IR_{HH}}\\ \beta=\dfrac{IR_{HH}}{IR_{NL}+IR_{HH}}\end{cases} \tag{8}$$

$IR_{NL}$ and $IR_{HH}$ in formula (6) can express as:

$$IR_{NL}=\frac{NUM_{L\_right}}{NUM_{L\_Total}} \tag{9}$$

 

$$IR_{HH} = \frac{NUM_{H\_right}}{NUM_{H\_Total}} \quad (10)$$

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 $O_H$.

# 4. Experimental Analysis and Discussion

## 4.1. Database and Experimental Setting

Experimental corpus base Mandarin Affective Speech Corpus (MASC) and Emotional Prosody Speech and Transcripts (EPST). MASC has 23 female and 45 male speakers' utterance in Chinese mandarin with 5 emotional classifications (neutral, angry, happy, scared, and sad classifications). Every speaker has 5 phrases and 60 sentences in every emotional state. Each phrase lasts 0.8 second averagely, while each sentence lasts 2 seconds averagely. Besides, there are 2 short passages with average duration of 15 seconds per passage in neutral state. EPST is the first emotional speech corpus released by Linguistic Data Consortium (LDC). It includes 8 actors (3 male, 5 female). 7 speakers of them provide their English speech in 14 emotional classifications and their neutral speech with different distance. The corpus used in the experiment were split into 3 parts: Speeches of the first 18 people (7 female and 11 male) in MASC were taken as development data to obtain fitting parameters; Speeches of the remains in MASC were test data. 2 short passages of every speaker were used to train speaker model, and the other 15,000 sentences were used as testing speeches; In addition, speeches of 7 speakers in EPST corresponding with 5 same emotional classifications as MASC were treated as extended test data. About 5 minutes neutral speeches of each speaker in normal distance were used to train speaker model. 5 kinds of emotional sentences with total count 670 were taken as testing speech.

In the experiment, 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. The weight coefficients α and β, baseband mapping function f, and gender models are all got from the dates in development data. The order of f is set According to bi-model approach base on equal weight. Take 11 as the order of male f, and 5 as the order of female f.

For verifying the validity of two kinds fusion weight estimating strategy based on the score reliability assessment, this part will compare the four methods of recognition performance on the MASC corpus and EPST corpus. The four methods are: the bi-model method fusion weight estimating strategy based on the score reliability assessment (score difference), the bi-model method based on the weight strategy of recognition rate (recognition rate), the bi-model method based on the equal weight (equal weight) and the traditional GMM-UBM method (datum).

## 4.2. Experimental Results on the MASC

**Table 1. The evaluation result on the MASC.**

| Method | IR (%) | | | |
|---|---|---|---|---|
| | Datum | Equal weight | Recognition rate | Score difference |
| Neutral | 96.23 | 95.40 | 95.47 | 95.30 |
| Angry | 31.50 | 36.37 | 37.93 | 38.43 |
| Happy | 33.57 | 37.57 | 39.17 | 39.80 |
| Scared | 35.00 | 37.10 | 38.97 | 39.27 |
| Sad | 61.43 | 60.70 | 60.93 | 61.10 |
| Average | 51.55 | 53.43 | 54.49 | 54.78 |

Experimental results with four methods on MASC corpus were shown in **Table 1**. Relative to the datum, basing on 3 different weight estimating strategy recognition rates of high differences emotions testing speech are improved obviously（4.87% - 6.93% on angry speech, 4.00% - 6.23% on happy speech and 2.10% - 4.27% on scared speech). And the performance of the low different emotional testing speech has declined slightly (0.76% - 0.93% on neutral speech and 0.33% - 0.73% on sad speech). For the bi-model method, these two weight estimating strategy is better the equal weight strategy, especially the recognition property of three high differences emotional testing speech has a obviously improvement. This improvement mainly benefit from that two assessment methods can assess the score reliability effectively, so that we can merge two different score effectively. We still find that the recognition property of the system which is on the low differences emotional speech has declined slightly. This problem is mainly caused by inaccurate of mis-match testing. It is known easily that it will have some negative impact if we use low mis-match parts to score on the virtue high differences emotional model. Despite these shortcomings, the two weight estimating strategy proposed in our paper are increased 3.23% compared with standard on the recognition rate, and it is increased 1.35% than the traditional method.

 

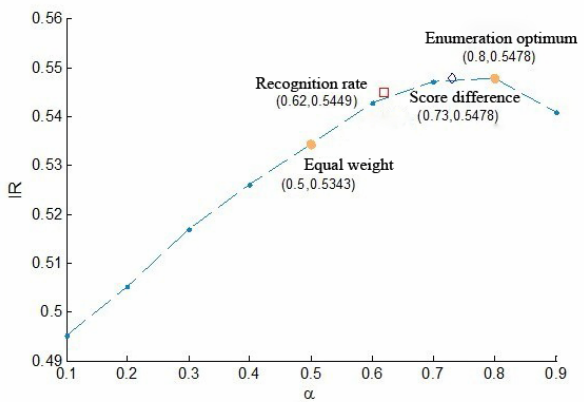


**Figure 2. The IR on MASC for BESR method with various weights.**

For checking the effectiveness of the weight estimating strategy, this part will compare the different weight coefficient on the bi-model method which the recognition performance of the MASC corpus. **Figure 2** count the recognition rate of bi-model method which is based on $\alpha = \{0.1,0.2,\cdots,0.9\}$ , $\alpha = 0.62$ (based on the weigh strategy of recognition rate) and $\alpha = 0.73$ (based on fusion weight strategy of scores average relative loss difference). We can know something from the figure that when $\alpha > \beta$ ($\alpha + \beta = 1, \alpha > 0.5$), the system performance is better than $\alpha \le \beta$, this tell us the score is more reliability which we use neutral model $Z_N$ to determine the low mis-match part (It is similar with low differences emotional speech $O_L$). The rational weigh α should greater than the score weigh $\beta$ of $Z_H$ on the high mis-match part. In addition, the recognition performance in the situation bi-model which is $\alpha = 0.73$ is same as the recognition performance of the best weigh ($\alpha = 0.8$) which is from enumerating $\alpha = \{0.1,0.2,\cdots,0.9\}$. And when $\alpha = 0.62$, the recognition performance of bi-model approximate with the best performance.

## 4.3. The Evaluation Result on the EPST

**Table 2. The evaluation result on the EPST.**

| Method | IR（%） | | | |
|---|---|---|---|---|
| | Datum | Equal weight | Recognition rate | Score difference |
| Neutral | 93.75 | 93.75 | 93.75 | 93.75 |
| Angry | 47.48 | 48.92 | 49.64 | 49.64 |
| Happy | 39.62 | 44.65 | 46.54 | 45.28 |
| Scared | 39.72 | 49.65 | 49.65 | 49.65 |
| Sad | 66.89 | 72.19 | 72.19 | 72.19 |
| Average | 53.88 | 58.66 | 59.25 | 58.96 |

In this paper, we can get the bi-model score weighting coefficients $\alpha$ and $\beta$ from MASC parameters development data by using our two Fusion Weight Estimating Strategies. Using the bi-model approach achieve good results in speaker recognition assessment experiment on MASC text set. EPST corpus has different culture background with MASC. We will verify the validity of the coefficient by bi-model approach in EPST corpus. Experimental results with four methods in EPST corpus were shown in **Table 2**. We can find that three bi-model approaches based on different score fusing strategies are also improved obviously than by baseline in performance of high differences emotion testing speech. In contras to MASC, recognition rate for sad speech is improved too. It is mainly because people with different culture also express his emotion very differently and pitch mean of sad speech in EPST corpus is noticeable higher than neutral speech. Because there is a big difference between two speech database, using $\alpha$ and $\beta$ get from the parameters developed dates in MASC corpus can't significantly improve the recognition rate of bi-model approach, but it is still effective. The overall recognition rate from EPST corpus, bi-model approach based on two score reliability assessment strategies is also improved 0.30% and 0.59% than equal weight. In addition, we can also get from **figure 3**, the recognition rate in EPST corpus with bi-model approach, the fusion weight strategy based on recognition rate ($\alpha = 0.62$) is better than by enumerate strategy ($\alpha = 0.8$), and the fusion weight strategy based on scores average relative loss difference $\alpha = 0.73$ also approach to optimal enumerate ($\alpha = 0.8$).

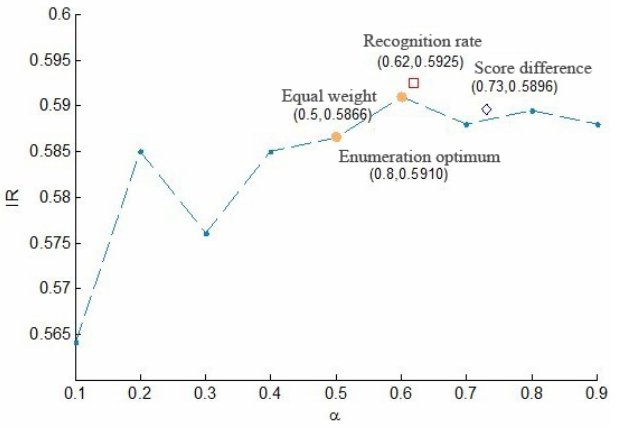


**Figure 3. The IR on EPST for BESR method with various weights.**

## 5. Conclusions

It is clearly unreasonable to fuse the score equal weight in Bi-model. So in this paper, we propose two fusion weight strategies based on score reliability fusion: by

 

utilizing identification rate and scores average relative loss difference. Systemic risk caused by Unreliability of Virtual synthesized speech could be reduced by using our strategy. In MASC, bi-model based on score reliability fusion is improved 3.23% than by baseline, and 1.35% than bi-model based on equal weight. In extended test of EPST corpus, using our new strategy is also improved 0.59% than before.

## 6. Acknowledgements

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] M. Liberman, *et al*., "Emotional Prosody Speech and Transcripts," Philadelphia; Linguistic Data Consortium. 2002.

[2] T. Huang and Y. Yang, "Learning Virtual HD Model for Bi-model Emotional Speaker Recognition," ICPR, Istanbul, Turkey, 23-26 Aug. 2010, pp. 1614-1617.

[3] D. G. Childers, J. J. Yea and E. L. Bocchieri, "Source/ Vocal-tract Interaction in Speech and Singing Synthesis," *Proc Stockholm Music Acoust Conf*, 1983, pp. 125-141.

[4] D. G. Childers and C. F. Wong, "Measuring and Modeling Vocal Source-Tract Interaction, *Ieee Transactions on Biomedical Engineering*, Vol. 41, No. 7, 1994, pp. 663-671. doi:10.1109/10.301733

[5] H. Akaike, "A New Look at the Statistical Model Identification," *Automatic Control, IEEE Transactions on*, Vol. 19, No. 6, 1974, pp. 716-723. doi:10.1109/TAC.1974.1100705

[6] L. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection, Acoustics,Speech and Signal Processing," *IEEE Transactions on*, Vol. 25, No. 1, 1977, pp. 24-33. doi:10.1109/TASSP.1977.1162905

[7] L. H. Cai, D. Z. Huang and R. Cai, "Basis of Modern Speech Technology and Application," Tsinghua University press, Beijing, China, 2003.

 