International Journal of Control and Automation
Vol. 7, No. 12 (2014), pp. 339-352
http://dx.doi.org/10.14257/ijca.2014.7.12.31

A Continuous Abnormal Speech Detection Method Based on Time Domain features Weighted

He Jun[1], Ji-chen Yang[2], Qing-hua Zhang[1], Guo-xi Sun[1] and Jian-bing Xiong[1]

[1] Guangdong Province Key Lab of Petrochemical Equipment Fault Diagnosis Guangdong University Petrochemical Technology Maoming City, Guangdong province, China

[2] South China University of Technology School of Electronic and Information Engineering, Guangzhou, China

E-mail:hejun_723@126.com; NisonYoung@163.com ; fengliangren@tom.com; sguoxi@126.com; xiongjianbin@21cn.com;

Abstract

In this brief, a novel pathological continuous speech detection method based on time domain features weighted. First, different optimal threshold for time domain features, including zero crossing ratio, short-time energy and autocorrelation, are obtained from training speech data. Second, a difference evaluation technique is proposed, and with it, the difference of the same time domain feature selected from testing speech data and training speech data were obtained. Finally, to distinguish a given speech well, a novel weighting method based on difference evaluation for each kinds of time domain is employed, respectively. Experiments were conducted on the pathological speech database to prove the power and effectiveness of the proposed method. Results obtained shown that this method outperforms other early proposed time domain feature method, creating a more reliable technique for pathological continuous speech detection.

Keywords: *pathological speech, continuous speech, time domain feature, weighting factor*

1. Introduction

The robustness of speech recognition has been paid attention to and devoted to by many scholars in speech signal processing, and so some encouraging progresses have been obtained in recently years [1-2]. However, the performance of many speaker recognition methods has a sharp dropped, when its face the abnormal speech which produced under the vocal organ cannot work well [3-4]. In order to overcome the drawbacks of these method, many researchers, devoting to promote the application of voice products, do their best working on the abnormal speech signal processing, and some reports have been published [5-6]. And then, in recently years, a special column about abnormal speech signal processing has been opened in European speech conferences, some abnormal speech database were introduced and used, which reported comprehensively [7-8]. With further research, the abnormal speech was divided into three types: 1) disguise voice, pronouncing by the speakers who change the ways of them speaking intentionally, such as chewing voice, knead nasal voice, imitative voice, and so [9-10]; 2) pathological voice, pronouncing by the speakers with whose vocal organ cannot work well, such as cold voice, drunk voice, running voice, G-force voice, and so on [5, 11-12]; 3) emotion voice, generating under a uncontrolled emotion, such as creaming voice, weeping voice, and so on [12-14].

A promising research field in speech signal processing, disguise voice, was very useful in the judicial help to identify the threats speech came from which speaker [15-16] . However, in

ISSN: 2005-4297 IJCA

the actual application of the products based on speech, in generally speaking, the users can meet all the requirements to help them to use the equipment smoothing. So the users do not change their pronunciations means especially when they are using the products. Emotion speech research has a great help for analyzing the emotion which masked in the speech of speaker. However, in some entrance guard systems based voiceprint, in order to facilitate pass through the authentication by systems, the speaker can take a few minutes to calm their emotions.

Now today, the environmental polluted more and more serious and the pressure of life is more and bigger, which make it is easy to cause vocal organs lesions, and it was the root cause that aroused the speakers' voice change. These kinds of abnormal speech have a long duration and the factors aroused the speech change are complex. This was a key problem for any speaker recognition systems based on voiceprint needed be solved. The robust of speaker recognition systems was focused on how to deal with the abnormal speech. So it became a new hot topic in the field of speaker recognition even in the speech signal processing fields, rapidly. To improve the ability of responding to abnormal voices for speaker recognition system, firstly, the system must have the ability to distinguish the abnormal voice and normal voice, which was the basis problem of speaker recognition.

Currently, most researches about abnormal speech were focused on biomedical engineering, aiming to detect the pathological speech from normal speech, in the other words, in order to find lesions in the vocal organs early through analysis the speech quality [5, 11, 17]. Recently, pathological voice detection and tracking have achieved fast development; some pathological voice detection methods were published. These methods for pathological voices detection can be divided into the following four types mainly; 1) pathological voice detection methods based on perturbation [18], such as Absolute Jitter (AJ), Relative Average Perturbation (RAP), Amplitude Perturbation (AP), Shimmer Percent, Absolute Shimmer (AS), *etc.*, 2) detection methods based on noise, such as Harmonics to Noise Ratio (HNR)[19], Normalized Noise Energy (NNE), Glottal to Noise Excitation Ratio (GNER)[20], Sub-Harmonic to Noise Ratio (SHNR), *etc.,* 3) based on model, such as based on GMM or based on GMM-SVM [11], *etc.,* 4) based on nonlinear method, such as based LLE [21], *etc..,*

From what has been discussed above, sustained vowels detection was focused on in most of the reported. However, the work on continuous pathological speech detection was reported rarely. Although sustained vowels offer a more controlled way of quantifying voice characteristics, and in general produce good classification results, because of the sustained vowels can characterize the cases about vibration of vocal cords in the period of pronunciation. However, the sustained vowels can't representative of speaker as they are of discoursing. On the other hand, continuous speech signals capture important attributes of speaker information. At the same time, the methods mentioned above had high computational complexity and was not suitable for online detection.

With this background, the goal of the work presented in this paper is to evaluate time-domain features and with weighting them to distinguish continuous normal speech and pathological speech. With an emphasis on simplicity and effectiveness of the proposed method, we weighted different time domain feature through evaluating the degree of variation for time domain feature of normal speech and abnormal speech. The overview of the paper is as follows, Section 2 briefly reviews time domain features for speech. The design of time domain features weighting is presented in Section 3. Discussion and results analysis with this proposed and the previous work are presented in Section 4. This paper concludes in Section 5.

2. Time Domain Features Introduction

2.1. Short-Time Energy (STE)

The short time energy is the energy of short speech segment. Short time energy is a simple and effective classifying parameter of voiced and unvoiced segments. Energy is also used for detecting end points of utterance and quality of speech for speaker. The long term definition of signal energy is as below[22]:

$$E = \sum_{m=0}^{\infty} x^2(m) \tag{1}$$

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) = x^2(n-N+1) + \cdots + x^2(n) \tag{2}$$

For a short-time speech signal, an nth frame window is applied on this signal:

$$x_n(m) = x(m)w(n-m), n-N+1 \le m \le n \tag{3}$$

where $n = 0, 1T, 2T, \cdots, N$ denotes the length of the window and T is the frame-shift. In where the high energy would be classified as voiced and lower energy as unvoiced. Because of the above property, in this paper the short-time energy used to characterize the vibration of vocal cords under the cold. As the vocal cords open as usually, but it can't open incompletely, and when closing it can't close well, in this case the short-time energy were fluctuated heavily. Figure 1 (a) and figure 1 (b) given the STE for speech template as below:

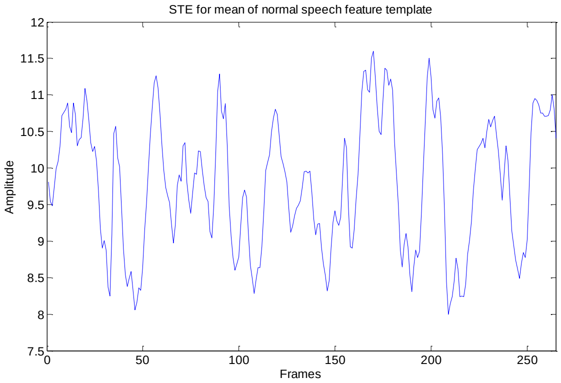

Figure 1. (a) STE for Mean of Normal Speech Feature Template

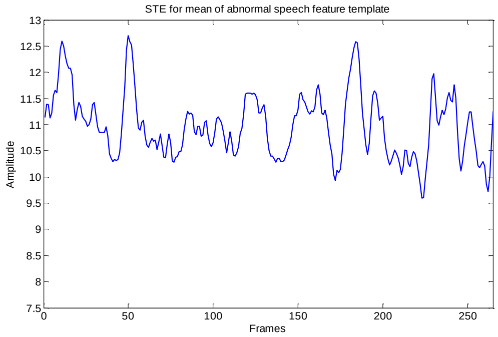

Figure 1. (b) STE for Mean of Abnormal Speech Feature Template

Figure 1 shows: the STE of abnormal speech feature template from Figure 1 (a) were concentrated in 10 and 11.5, which is bigger than Figure 1 (b) given.

2.2. Zero Crossing Rate (ZCR)

The Zero Crossing rate is used for counting the number of zero crossing in a giving speech. Voiced speech segments have low ZCR as compare to ZCR of unvoiced segments.

In the context of discrete-time speech signals, a zero crossing is said to occur if successive sample have a different algebraic signs. ZCR is measurement of “frequency composition” of a signal; this is more valid for narrowband signals such as sinusoids.

The rate at which zero crossing occurs is a simple measure of the frequency content of a signal.

A definition for zero crossing rates is given as following:

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\text{sgn}[x(m)] - \text{sgn}[x(m-1)]\right| w(n-m) \tag{4}$$

where

$$\text{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases} \tag{5}$$

and

$$w(n) = \begin{cases} \frac{1}{2N}, 0 \leq n \leq N-1 \\ 0, \quad otherwise \end{cases} \tag{6}$$

The model for speech production suggests that the energy of voiced speech is concentrated below about 3Hz because of the spectrum fall of introduced by the glottal wave, whereas for unvoiced speech, most of the energy is found at higher frequencies. Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that the zero-crossing rate is high; the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced. The ZCR of mean of normal speech and abnormal speech feature template given as Figure 2 (a) and Figure 2 (b), respectively.

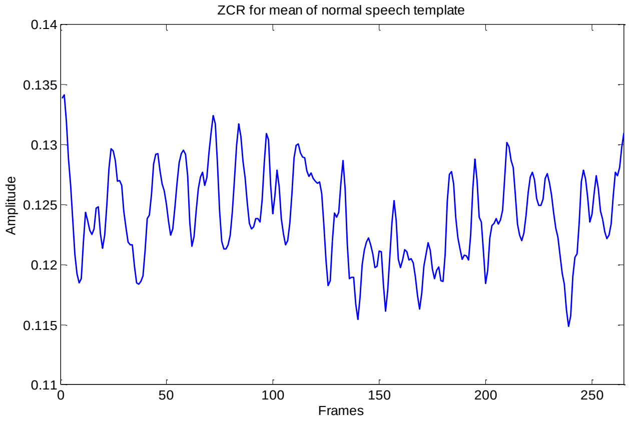

Figure 2. (a) ZCR of Mean of Normal Speech Feature Template

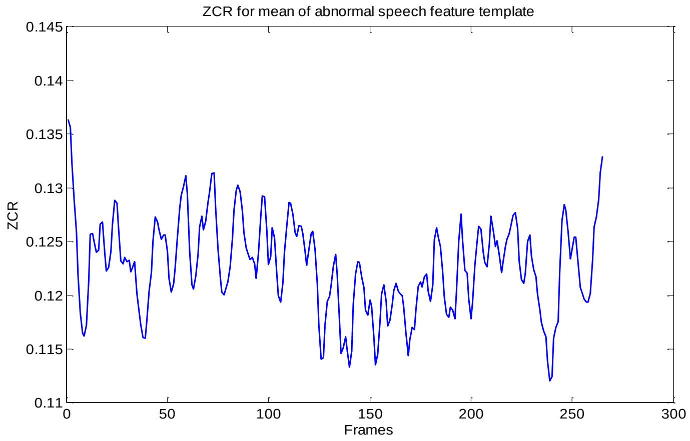

Figure 2. (b) ZCR of Mean of Abnormal Speech Feature Template

2.3. Short-Time Auto Correlation (STAC)

Short-time auto correlation is an important method to get the pitch for dullness, which was defined as the following:

$$R_n(k) = \sum_{m=0}^{N-1-k} [x(n+m)w'(m)][x(n+m+k)w'(m+k)] \quad (7)$$

However, when calculating the short-time auto correlation, the finite window length for the selected voice segment was N, and the summation of the ceiling was $N-1-k$. While the data used to calculate is decreasing which leading to the amplitude of the correlation function decrease with the increase of k. To solve this problem, a modified speech short-time autocorrelation was defined as the following:

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w_1(n-m)x(m+k)w_2(n-m-k) \quad (8)$$

Assuming that $m = n + m'$, and then eq(8) redefined as:

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(n+m')w_1(-m')x(n+m'+k)w_2(-m'-k) \quad (9)$$

Defined that:

$$\begin{cases} \hat{w}_1(m) = w_1(-m) \\ \hat{w}_2(m) = w_2(-m) \end{cases} \quad (10)$$

and then

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} x(n+m)\hat{w}_1(m)x(n+m+k)\hat{w}_2(m+k) \quad (11)$$

And

$$\begin{cases} \hat{w}_1(m) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & other \end{cases} \\ \hat{w}_2(m) = \begin{cases} 1, 0 \le n \le N-1+k \\ 0, \quad other \end{cases} \end{cases} \quad (12)$$

When meeting the condition $m+k \le N-1+k$, the equation $\hat{w}_2(m+k) \ne 0$ permanent establishment. So, equation (11) rewritten as:

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k) \tag{13}$$

The detailed STAC distributed of each frame given as following figure.

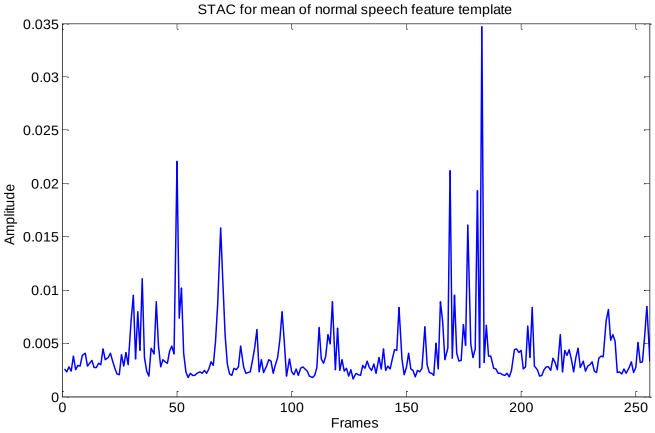

Figure 3. (a) STAC for Mean of Normal Speech Feature Template

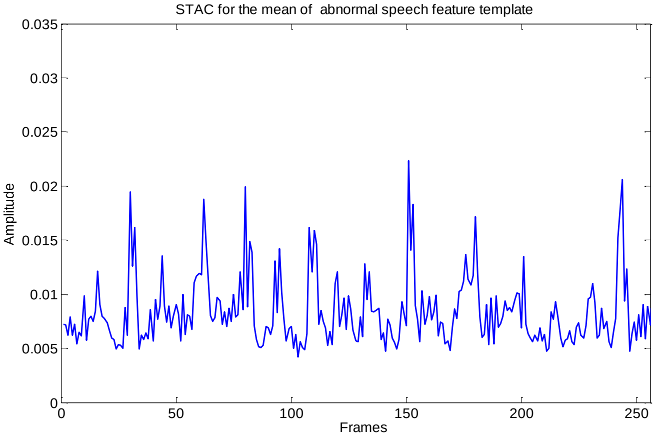

Figure 3. (b) STAC for Mean of Abnormal Speech Feature Template

2.4. Short-Time Magnitude (STM)

Given a time serial speech signal $x(t)$, $x_n(m)$ denote the nth frame signal through framed and windowed processing for $x(t)$, $x_n(m)$ can be given as follow:

$$x_n(m) = \omega(m)x(n+m) \qquad 0 \le m \le N-1 \tag{14}$$

where $n = 1, 1T, 2T, \cdots, N$, and N denote the length of frame for speech signal $x(t)$, T was the length of frame shift, $\omega(m)$ was the window function. Then, the short-time energy of nth frame given as below:

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{15}$$

E_n, was a function which used to measure the diversification of magnitude for speech signal, has an obvious drawback that it sensitive to high-level signal. So the other method used to measure the magnitude, namely, short-time average magnitude function, which has given as follow:

$$M_n = \sum_{m=0}^{N-1} | x_n(m) | \tag{16}$$

The difference between M_n and E_n is neither small sample nor big sample did not bring bigger which aroused by square operation.

STM for mean of normal speech and abnormal speech feature template given as Figure 4 (a) and Figure 4 (b).

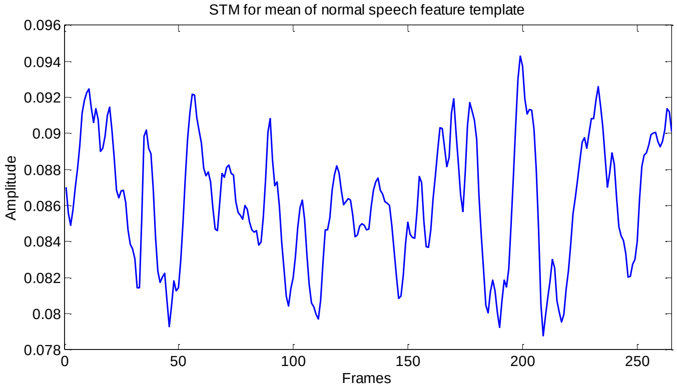

Figure 4 (a). STM for Mean of Normal Speech Feature Template

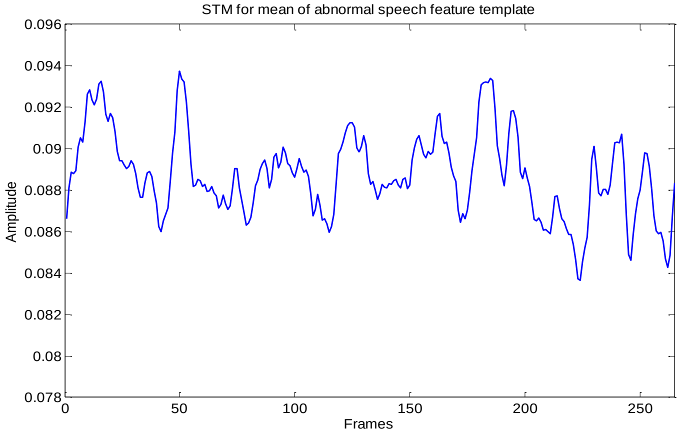

Figure 4. (b) STM for Mean of Abnormal Speech Feature Template

The Figure 4 (a) and Figure 4 (b) illustrated the STM value of abnormal speech gathered between 0.085 and 0.094.

3. **Algorithm Description**

Because of feature extracting simple, convenient and with low computational cost, time-domain features were used in many online speech processing systems. Due to the vocal organs lesions, the utterances from the speaker can arouse abnormal, the performance of speech recognition method based on time-domain feature dropped sharply. In this section, we will describe the process about time-domain features weighting. The details of this proposed as the following Figure 5:

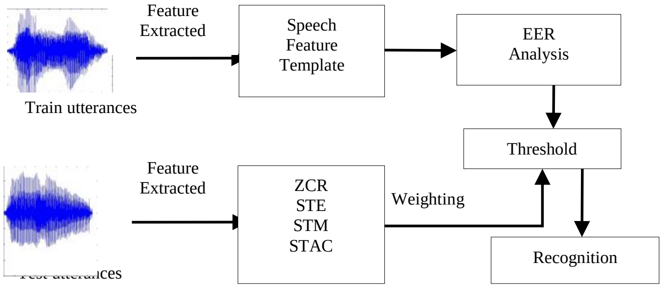

Figure 5. The Block of this Proposed

3.1. Time-domain Feature Template Building

Assuming that U_i denoted that *ith* given utterance, which can be divided into N frames, the time-domain features extracted for *ith* given utterance can be defined as:

$$f_t^i = \{ f_{1,t}^i, \cdots, f_{k,t}^i \cdots, f_{n,t}^i \}, t = 1,2,3,4; n = 1, \cdots N \tag{17}$$

where t used to represent the type of time-domain feature, such as f_t^i denotes the ZCR feature of *ith* given utterance.

Mass training normal speech used to obtain the speech feature template for them, and then a speech feature template about normal speech can be obtained which organized as follows:

$$\boldsymbol{F}_{\boldsymbol{NST}}^{\boldsymbol{t}} = \begin{bmatrix} f_{1,t}^1 & f_{2,t}^1 & \cdots & f_{n,t}^1 \\ f_{1,t}^2 & f_{2,t}^2 & & f_{n,t}^2 \\ \vdots & \vdots & & \vdots \\ f_{1,t}^m & f_{2,t}^m & & f_{n,t}^m \end{bmatrix} \tag{18}$$

In equation (18), $\boldsymbol{F}_{\boldsymbol{AST}}^{\boldsymbol{t}}$ was obtained by using the same method mentioned above to abnormal speech.

3.2. Threshold Obtained by EER Analysis

In this section, the equate error ratio (EER) method was used to analysis the speech template and obtained the optimal characteristics threshold for distinguish between normal and pathological speech. EER was one of the common used performance evaluation indicators for all kinds of speaker recognition system or speech recognition system widely.

In order to carry out the EER analysis for speech feature template well, we deal with the speech feature template as following:

$$\overline{f_t^i} = \frac{1}{N}\sum_{k=1}^{\mathrm{N}} f_{k,t}^i \tag{19}$$

where $\overline{f_t^i}$ denotes that the mean value of *ith* given utterance with the time-domain feature type t , $t \in (ste, zcr, stac, sta)$, and then the speech feature template can be rewritten as follows:

$$\boldsymbol{F'^{t}_{SFT}} = [\overline{f_t^1}, \overline{f_t^2}, \cdots, \overline{f_t^m}]^T \tag{20}$$

Assuming that $\boldsymbol{F'^{t}_{p\text{-}SFT}}$ and $\boldsymbol{F'^{t}_{n\text{-}SFT}}$ illustrated the time-domain feature template for pathological speech and normal speech, respectively. Before carrying out EER analysis, the array of $\boldsymbol{F'^{t}_{p\text{-}SFT}}$ and $\boldsymbol{F'^{t}_{n\text{-}SFT}}$ were sorted by ascending, respectively. And the EER analysis carried out as the figure5 illustrate.

$$\begin{array}{l} \mathrm{F}^{'t}_{\mathrm{n\_SFT}} \rightarrow \overbrace{v_{n,1} \quad v_{n,2} \quad \cdots \quad \cdots \quad v_{n,r} \quad \cdots\cdots\cdots \quad v_{n,m2}}^{m2} \\ \hphantom{\mathrm{F}^{'t}_{\mathrm{n\_SFT}} \rightarrow v_{n,1} \quad v_{n,2} \quad \cdots \quad \cdots \quad} \underbrace{v_{p,1} \quad \cdots\cdots\cdots \quad v_{p,k} \quad \cdots \quad \cdots \quad v_{p,m1}}_{m1} \leftarrow \mathrm{F}^{'t}_{\mathrm{p\_SFT}} \end{array}$$

Fig.5 illustrate for fast EER analysis

In the Fig.5, $v_{n,r}$ was selected as the initial threshold at the beginning of EER analysis, then moving the threshold in the direction of $v_{n,m2}$ step by step. In all the EER analysis process, an optimal threshold can be obtained which made the EER minimize.

Through the EER analysis, we can obtain a threshold sign function as below:

$$sgn(\lambda_i - \lambda_j) = \begin{cases} 1, & \lambda_i - \lambda_j > 0 \\ -1, & \text{otherwise} \end{cases} \tag{21}$$

The equation (21) used to weight the time domain features extracted from given testing utterance,

3.3. Weighting Factor

Given a testing utterance, its time-domain feature denoted by f_t , and then the two distances of f_t and the time-domain feature template of normal speech and abnormal speech calculated by the following equation, respectively:

$$d_{t,n} = \sum_{i=1}^{N} | f_t^i - F_{NST}^t | \tag{22}$$

$$d_{t,a} = \sum_{i=1}^{N} | f_t^i - F_{AST}^t | \tag{23}$$

where $d_{t,n}$ and $d_{t,a}$ denotes the distance for testing utterance and F_{NST}^t and F_{AST}^t , respectively. Then a comparing the size of $d_{t,n}$ and $d_{t,a}$, if it meets the condition $d_{t,n} < d_{t,a}$, a hypothesis that the testing utterance judged as normal speech was obtained. Otherwise it will be judged as abnormal speech. To test the hypothesis, a weighting factor calculating method is proposed. The detailed description of the method as below:

$$w_t = 1 + \mathrm{sgn}(d_{t,n} - d_{t,a}) \times \frac{|\, d_{t,n} - d_{t,a} \,|}{d_{t,a}} \tag{24}$$

where w_t denotes the weighting factor of a given test utterance, then the weighting factor used to weight time-domain feature for the test utterance as the follows:

$$F_{nt} = w_t \times F_{ot} \tag{25}$$

where F_{nt} and F_{ot} denotes the new feature and original feature of the given test utterance, and then, the F_{nt} used to distinguish the test utterance.

4. Experiment and Analysis

4.1. Experimental Setup

998 normal utterances from 9 speakers (100 for each speaker or so) and 2508 abnormal utterances from 13 speakers, (193 for each speaker or so) which were text-independent continuous speech, selected from the PANSD [7] as the experimental speech data for this paper. All the experiment data were divided into two groups randomly as the ratio for each scheme shown in the Table I.

Table I. The Ratio and Segments for Each Scheme

Scheme	Ratio	Train		Test	
		n-num	p-num	n-num	p-num
1	3:7	272	746	726	1762
2	5:5	496	1250	502	1258
3	7:3	697	1730	301	778

Illustrated from the Table I, the n-num and p-num denote the number of segment for normal speech and pathological speech for training and testing, respectively. The ratio denotes the ratio of the n-num of train segments and that of test segments.

All speech material used for doing experiment is mono channel WAV format file, which digitized at 16bits, at 16 kHz sampling frequency. All speech material are cut into 4s by given program automatically, with removing the entirely silence component by artificial processing. Each segment is firstly divided into frames by 30ms duration of each frame with 50% overlap.

4.2. Results and Analysis

In this paper, the total recognition ratio used to evaluate the performance of each time domain feature method, giving as the below:

$$r_{total} = \frac{R_n + R_p}{N_N + N_P} \times 100\% \tag{26}$$

where R_n and R_p denote the number of normal utterance judged as normal and the number of abnormal utterance judged as abnormal utterance, respectively. N_n and N_p denote the total normal utterances and abnormal utterance for recognition, respectively.

According to the proportion mentioned in Table I, we obtain the results as following. The Table II given the results for scheme I, the Table III and Table IV show the results for scheme II and scheme III, respectively.

Table II. Recognition Ratio for Mentioned Time Domain Method and with its Weighted in Scheme I

Time-domain	threshold	r_{total} (%) Un-weighted	r_{total} (%) With weighted
ZCR	0.1235	55.16	61.87
STE	6.2828	53.22	56.31
STM	0.0827	54.71	58.43
STAC	0.0058	52.88	53.79

Table III. Recognition Ratio for Mentioned Time Domain Method and with its Weighted in Scheme II

Time-domain	threshold	r_{total} (%) Un-weighted	r_{total} (%) With weighted
ZCR	0.1214	56.82	62.61
STE	5.4093	53.72	57.18
STM	0.0767	55.90	58.70
STAC	0.0053	53.72	55.05

Table IV. Recognition Ratio for Mentioned Time Domain Method and with its Weighted in Scheme III

Time-domain	threshold	r_{total} (%) Un-weighted	r_{total} (%) With weighted
ZCR	0.1232	57.91	63.38
STE	4.9917	53.35	58.86
STM	0.0735	55.94	59.04
STAC	0.0049	54.58	56.17

From Table II, Table III and Table IV show: the weighting method proposed in this paper, which can improve the performance of various method based time-domain characteristics for continuous speech recognition. As the training data increasing, the threshold values for various time-domain characteristics turned more accurate and the performance of the weighting algorithm stable gradually. And the performance of ZCR affected smaller with the amount of training data, it suited for the continuous speech recognition compared with the other methods mentioned in this paper. However, due to the speakers have a cold, which aroused their vocal organ functional disorders. The mainly manifestation is the glottis cannot work well, such as it open incompletely or it is not closed tightly. Resulting to generate a hoarse noise in the speaking, this can affect the ZCR and STE of the utterance apparently.

From Table II, Table III and Table IV show: it is a long way to go for recognizing the pathological continuous speech and normal speech, which impeding the process of speech based products market.

5. Conclusion

In this paper, a new algorithm for pathological speech detection based on time-domain feature weighting, is proposed. Four time-domain methods used to recognize the pathological speech was introduced in this paper, the performance of each time-domain feature has improved by weighting their time-domain feature, this kinds of method based on time-domain weighting were suitable for online speech detection system which can satisfy the command of low-complexity.

However, the results show that there is a long way to go in the field of pathological continuous text-independent speech detection. Although, a little progress we have made, however, the time-domain characters of continuous speech with uncertainly context, have huge difference. So the performance of this proposed does not meet the actual requirements for pathological continuous speech detection, and then methods of better performance are the direction of our efforts.

Acknowledgements

This work was supported by the National Nature Science Foundation of China (No.61301300), China postdoctoral Science Foundation (No.2013M531850) and the Fundamental Research Funds for the Central Universities, SCUT, China (No.2013Zm0097), Open fund of Guangdong Province Key Lab of Petrochemical Equipment Fault Diagnosis.

References

[1]. N. Karamangala and R. Kumaraswamy, "Speaker Recognition in Uncontrolled Environment: A Review", Journal of Intelligent Systems, vol. 22, no. 1, (**2013**), pp. 49-65.

[2]. T. May, S. van de Par and A. Kohlrausch, "Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, (**2012**), pp.108-121.

[3]. C. W. Maina and J. M. Walsh, "Compensating for noise and mismatch in speaker verification systems using approximate Bayesian inference", Information Sciences and Systems, 2011 45th Annual Conference on, pp. 1-6.

[4]. T. May, S. van de Par and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling", Audio, Speech, and Language Processing, IEEE Transactions on, no. 99, (**2011**), 1-1.

[5]. R. Tavares, N. Brunet, S. C. Costa, S. Correia, B. G. Aguiar Neto and J. M. Fechine, "Combining entropy measurements and cepstral analysis for pathological voice assessment", Biosignals and Biorobotics Conference, ISSNIP, (**2011**), pp. 1-5.

[6]. W. Ning, P. C. Ching, Z. Nengheng and Lee Tan, "Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 1, (**2011**), pp. 196-205.

[7]. H. J. L. Y-X. He and Q-H. L. Wei, "Speaker Recognition Algorithm for Abnormal Speech Based on Abnormal Feature Weighting", Journal of South China University of Technology (Natural Science Edition), vol. 3,(**2012**), pp. 106-111.

[8]. http://www.masseyeandear.org/.

[9]. C. Zhang and T. Tan, "Voice disguise and automatic speaker recognition", Forensic Science International, vol. 175, no. 2–3,(**2008**) pp. 118-122.

[10]. P. Perrot, G. Aversano and G. Chollet, "Voice Disguise and Automatic Detection: Review and Perspectives" Nonlinear Speech Processing Springer Berlin / Heidelberg, vol. 4391, (**2007**), pp. 101-117.

[11]. X. Wang, J. Zhang and Y. Yan, "Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach", Journal of Voice, vol. 25, no. 1, (**2013**), pp. 38-43.

[12]. R. Martsyshyn and Y. Rashkevych, "Technology of the speaker verification under stress". CAD Systems in Microelectronics , 2011 11th International Conference The Experience of Designing and Application of, pp. 438-439.

[13]. Y. Qin and Y. Qi. "EEMD-Based Speaker Automatic Emotional Recognition in Chinese Mandarin", Appl. Math, vol. 8, no. 2, (**2014**) pp. 617-624.

[14]. S. Zhang, L. Li and Z. Zhao, "Spoken emotion recognition using kernel discriminant locally linear embedding", Electronics Letters, vol.46, no. 19, (**2010**), pp. 1344-1346.

[15]. L. A. Khan, F. Iqbal and M. S. Baig, "Speaker verification from partially encrypted compressed speech for forensic investigation", Digital Investigation, vol. 7, no. 1-2, (**2010**), pp. 74-80.

[16]. P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence", Computer Speech &amp; Language, vol. 20, no. 2–3, (**2006**), pp. 159-191.

[17].L. Arias, x00F, J. D. o, J. I. Godino-Llorente, Sa, x, Lecho enz, N. n, V. Osma-Ruiz, D. Castellanos and G. Nguez, "Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients", Biomedical Engineering, IEEE Transactions on, vol. 58, no. 2, (**2011**), pp. 370-379.

[18].M. Brockmann, M. J. Drinnan, C. Storck and P. N. Carding, "Reliable Jitter and Shimmer Measurements in Voice Clinics: The Relevance of Vowel, Gender, Vocal Intensity, and Fundamental Frequency Effects in a Typical Clinical Task", Journal of Voice, vol. 25, no. 1, (**2011**), pp. 44-53.

[19].P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonic to noise ratio of a sample sound", Institute of Phonetic Sciences,University of Amsterdam, , vol. 17, (**1993)**, pp.97-110.

[20].D. Michaelis, T. Gramss and H. W. Strube, "Glottal to noise excitation ratio-A new measure for describe pathological voices", Acta Acustica United with Acustica, , vol.83, (**1997**), pp.700-706.

[21].J. Quansheng, L. Jiayun and J. Minping, "New method of fault feature extraction based on supervised LLE", Control and Decision Conference,Chinese, (**2010**)pp. 1727-1731.

[22].M. Jalil, F.A. Butt and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals", Technological Advances in Electrical, Electronics and Computer Engineering , International Conference on, IEEE, (**2013**) pp. 208-212.

Authors

Jun He, He was born in Shaoyang, Hunan province China in July, 1978. He received the B.Eng. degree in computer and application from National University of Defense Technology, Changsha, Hunan province, China, in 2002; he received the M. Eng. Degree in computer software and theory from South China Normal University, Guangzhou, Guangdong province, China, in 2008. He received Ph.D. in signal and information processing from South China University of Technology, Guangzhou, Guangdong province, China, in 2012.

Now he is working in Guangdong University of Petrochemical Technology and Guangdong key laboratory of petrochemical equipment fault diagnosis. His currents interest is abnormal speech detection, speaker recognition for abnormal utterance and large vibration machinery fault diagnosis, image processing, pattern recognition, machine learning.

Ji-chen Yang, He was born in Jieshou, Anhui province China in June, 1980. He received the B.Eng. degree in electronic and information engineering from Guangdong University of Petrochemical Technology, Maoming, Guangdong province, China, in 2004, he received the M. Eng. Degree in system engineering from Guangdong University of Technology, Guangzhou, Guangdong, China, in 2007.he received the Ph.D. in Telecommunication and information system from South China University of Technology, Guangzhou, Guangdong, China, in 2010.

Now he is a post doc. researcher in South China University of Technology. His currents interest is movie audio signal processing.

Qing-hua Zhang, He received the B.Sc. degree in Electrical Automation from Henan polytechnic university, China, in 1985 and the M.Sc. degree in Industrial Automation from South China University of technology, China, in 1995 and the Ph.D. degree in Control Theory and Control Engineering from South China University of Technology, China, in 2004. Now he is a professor and currently the dean of the Guangdong province Petrochemical Equipment Fault Diagnosis (PEFD) key laboratory in Guangdong University of Petrochemical Technology, China. His research interests include: condition monitoring & fault diagnosis of rotating machinery, intelligence control, and applications of intelligent algorithms.

Guo-xi Sun, He was born in Youyi, Heilongjiang Province, China, in October, 1972. He received B.E degree in Control Technology and Instruments from Tianjin University, Tianjin, Jiangsu province, China, in 1992; he received the M. Eng. Degree in signal and information processing from South China University of Technology, Guangzhou, Guangdong province, China, in 1999; He received Ph.D. in Signals and Systems from South China University of Technology, Guangzhou, Guangdong province, China, in 2006.

Now he is work in Guangdong University of Petrochemical Technology and Guangdong key laboratory of petrochemical equipment fault diagnosis. His currents interest is abnormal speech detection, speaker recognition for abnormal utterance and large vibration machinery fault diagnosis, image processing, pattern recognition, machine learning.

Jian-bin Xiong, He was born in Shaoyang, Hunan province China, in July, 1976. He received B.E. and PH. D in Control Theory and Control Engineering from Guangdong University of Technology, Guangdong, China in 2006 and 2012, respectively. He is now working at the computer and information school, Guangdong University of Petrochemical Technology, China.

His current research interests include signal processing, image processing, information fusion, and computer applications.