Jingjiao LI[1],Dong AN[1],Dan ZHAO[2],Caoqun RONG[1],Shuang MA[1]

Northeastern University (1), China Medical University (2)

TEO-CFCC Characteristic Parameter Extraction Method for Speaker Recognition in Noisy Environments

***Abstract**. This paper proposes TEO-CFCC characteristic parameter extraction method. Signal phase matching is applied to eliminate speech noise on the basis of CFCC characteristic parameter, and then Teager energy operator is added to the acquisition of CFCC characteristic parameter. In this way TEO-CFCC characteristic parameter is obtained and the energy of speech becomes one of the characteristic parameters for speaker recognition. Experiment results show that the recognition accuracy can reach to 83.2% in a -5dB SNR of vehicle interior noise environment by using TEO-CFCC characteristic parameter.*

***Streszczenie.** W artykule przedstawiono metodę wyznaczania parametrów charakterystycznych filtru TEO-CFCC. Zastosowano tu dopasowywanie fazowe sygnału, dla eliminacji z mowy szumów oraz operator Teagera do wyrugowania parametrów. Badania eksperymentalne pokazuję, że dokładność rozpoznania głosu wynosi 83,2% przy -5dB SNR we wnętrzu pojazdu. (**Wyznaczanie parametru charakterystycznego dla filtru TEO-CFCC w rozpoznaniu głosu w zaszumionym środowisku**).*

Keywords: Auditory Transform, Phase Matching, Energy Operator, Speaker Recognition
Słowa kluczowe: transformacja słuchowa, dopasowanie fazowe, operator Teagera, rozpoznanie mówcy.

Introduction

Accurate extraction of the voice of characteristic parameters is the key step in the field which includes automatic word's identification [1], speech automatic segmentation [2] and speaker recognition [3]. At present, the speaker individual character characteristic parameters which is widely used is still Mel Cepstrum Coefficients (MFCC) [4], but MFCC characteristic parameter for speaker recognition declines sharply in low SNR environments [5]. Extracting a speaker individual character characteristic parameters which is both effective and anti-noise is still one of the main problems in speaker recognition study. The Cochlear Filter Cepstral Coefficients (CFCC), which is firstly proposed by Dr Peter Li who works in bell LABS in 2011, is the characteristic parameter used for speaker recognition [6]. CFCC which is the characteristic parameter based on the auditory transform (AT) [7].is different from MFCC which are based on fast Fourier transform (FFT). The accuracy of MFCC drops to 41.2% when the signal-to-noise ratio (SNR) of the input signal is 6dB in white noise and car noise environment, but CFCC still achieve an accuracy of 88.3%. All these shows that CFCC features have shown strong robustness and are better than the MFCC. But when the SNR of the input signal is -6dB under white noise, the accuracy of the CFCC features drops to 20%.

Considering the analysis above, this paper applied the signal phase matching into the CFCC characteristic parameter, and then Teager energy operator is added to the acquisition of it. Finally, this paper proposed speaker characteristic parameters extraction method based on phase matching and TEO-CFCC under noise environment, improved CFCC characteristic parameter.

1 CFCC Characteristic Parameter

1.1 The principle of auditory-based transform

The Fourier transform (FT) is the most popularly used transform to convert signals from the time domain to frequency domain. However, the advantages of Fourier transform in the treatment of the linear signal are restricted to the treatment of the nonlinear signal. Although the discrete Fourier transform and fast Fourier transform which are make up for the shortage of the Fourier transform perform better in dealing with nonlinear transform, it still can't meet the need of the speech signal processing.

Reference [7] presents that Peter Li first put forward the concept of auditory-based transform. The auditory-based transform imitated the principle of sense of hearing. He firstly definite a Cochlear Filter function by $\psi(t)\in\mathbf{L}^2(R)$, $\psi(t)$ are required by the conditions in equation (1)-(3):

$$\int_{-\infty}^{\infty}\psi(t)dt=0 \tag{1}$$

$$\int_{-\infty}^{\infty}\left|\psi(t)\right|^2 dt<\infty \tag{2}$$

$$\int_{-\infty}^{\infty}\frac{\left|\Psi(\omega)\right|^2}{\omega}d\omega=C \tag{3}$$

Where $0<C<\infty$ also

$$\Psi(\omega)=\int_{-\infty}^{\infty}\psi(t)e^{-j\omega t}d\omega \tag{4}$$

Let $f\left(t\right)$ be any square integrable function, the auditory-based transform for $f\left(t\right)$ can be defined as :

$$T\left(a,b\right)=\int_{-\infty}^{\infty}f\left(t\right)\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)dt \tag{5}$$

Where a and b is real, so the equation above is the same with:

$$T\left(a,b\right)=\int_{-\infty}^{\infty}f\left(t\right)\psi_{a,b}(t)dt \tag{6}$$

Where

$$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \tag{7}$$

Let equation (6) written as discrete expression:

$$T[a_i,b]=\sum_{n=0}^{N}f[n]\frac{1}{\sqrt{|a_i|}}\psi\left[\frac{n-b}{a_i}\right] \tag{8}$$

1.2 CFCC characteristic parameter extraction method

The auditory-based transform is a new method of nonlinear signal processing. It is the same as the Fourier transform (FT) since both can be used as a filter bank and complete to convert signals from the time domain to frequency domain. The auditory-based transform in the speaker characteristics extraction field has a good application when it is firstly proposed. Reference [6] proposed speaker's CFCC characteristic parameter extraction method based on the auditory-based transform. A typical Cochlear Filter function equation is defined by Peter Li as follows:

(9) $$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\left(\frac{t-b}{a}\right)^{\alpha}\exp\left[-2\pi f_L\beta\left(\frac{t-b}{a}\right)\right] \times\cos\left[2\pi f_L\left(\frac{t-b}{a}\right)+\theta\right]u(t-b)$$

Where $\alpha>0$ and $\beta>0$, the value of θ should be satisfied with equation (1). $u(t)$ is the unit step function. Factor *b* is a time shift real, and factor *a* is a scale or dilation variable. The value of *a* can be determined by the current filter central frequency f_c and the lowest central frequency f_L in the cochlear filter bank:

(10) $$a={f_L}/{f_c}$$

The range of *a* is $0<a\le 1$. Normally $\alpha=3,\beta=0.2$.

Let $f\left(t\right)$ be a speech signal, add the equation (9) to the equation (5), we can get $T\left(a,b\right)$ from $f\left(t\right)$ through AT transform. The cochlear filter bank is intended to emulate the impulse response. The inner hair cells transform the speech signal after the time-frequency transform to the electrical signal that can be analysed by the human brain. Peter Li used the equation (11)-(13) to emulate this process:

(11) $$h(a,b)=T(a,b)^2;\forall T(a,b)$$

(12) $$S(i,j)=\frac{1}{d}\sum_{b=l}^{l+d-1}h(i,b),l=1,L,2L,\cdots;\forall i,j$$

(13) $$y(i,j)=S(i,j)^{1/3}$$

In this equation, $d=\max\{3.5\tau_i,20\text{ms}\}$, τ_i is the period of the central frequency of the *ith* band, $L=10\text{ms}$. Finally, the discrete cosine transform (DCT) is applied to $y(i,j)$ and we can get the CFCC characteristic parameter.

2 TEO-CFCC Characteristic Parameter Extraction Method

A complete speech signal contains frequency information and energy information [8]. The energy is one of the most basic parameters in the speech signal and it represents the size of a frame of the speech signal. Even for the same content of the text, the speech signal's energy differs from different people in the same environment [9]. In addition, the short-term energy that found out by a frame of speech signal is a scalar value, which can express the time domain features of the speech. CFCC parameter is a characteristic of ear auditory perception .They both reflect the different characteristics of the speech signal. Speaker characteristic parameters combining with the two characteristics can represent the speaker's personality more. Reference [10] proposed Speech Feature Based on Teager Energy Operator(TEO) and Dyadic Wavelet Transform. Teager Energy Operator was successfully used in the characteristic parameters extraction and improved speech recognition performance. In this paper, Teager energy is used to represent the energy of the speech signal characteristics.

TEO is a nonlinear operator which can to enhance the signal the background noise as well as contribute to feature extraction. For signal sampling points $x(n)$, the discrete expression of TEO is:

(14) $$T[x(n)]=x^2(n)-x(n+1)x(n-1)$$

In the noise environment, assuming that the speech signal observed is the sum between pure speech signal $s(n)$ and nonzero mean additive noise $w(n)$:

(15) $$x(n)=s(n)+w(n)$$

If TEO is got when put $x(n)$ into equation (14) without any treatment, the accuracy of the results must be influence because of noise. In order to eliminate the influence caused by noise on speech signal, reference [10] put forward the energy estimate methods to eliminate noise. This energy estimate methods could estimate the TEO of the pronunciation efficiently in zero mean additive noise conditions, but it did not apply the nonzero mean additive noise conditions. In order to solve the problem, this paper introduces the signal phase matching into TEO. The principle of matched-signal phase method of three-sensor array is applied to eliminate the influence of TEO by nonzero mean additive noise. First change the equation (15) to the form of mode and phase:

(16) $$\left|X(j\omega)\right|e^{j\psi}=\left|S(j\omega)\right|e^{j\alpha}+\left|W(j\omega)\right|e^{j\varphi}$$

In equation (16), $\left|X(j\omega)\right|$, $\left|S(j\omega)\right|$ and $\left|W(j\omega)\right|$ is amplitude spectrum. ψ, α and φ is phase angle. They are all the functions of ω. As is shown in figure 1, three sensor array of line is used to receive signals, θ is the angle between the direction of signal and the direction of Linear array's normal. The frequency domain forms of the output signal from three sensors are as followed:

(17) $$\left|X_1(j\omega)\right|e^{j\psi_1}=\left|S(j\omega)\right|e^{j\alpha}+\left|W_1(j\omega)\right|e^{j\varphi_1}$$

(18) $$\left|X_2(j\omega)\right|e^{j\psi_2}=\left|S(j\omega)\right|e^{j(\alpha-\omega\tau)}+\left|W_2(j\omega)\right|e^{j\varphi_2}$$

(19) $$\left|X_3(j\omega)\right|e^{j\psi_2}=\left|S(j\omega)\right|e^{j(\alpha-2\omega\tau)}+\left|W_3(j\omega)\right|e^{j\varphi_3}$$

Where $\tau=\frac{d}{c}\sin\theta$, d is array element spacing. c is the transmission speed of wave.

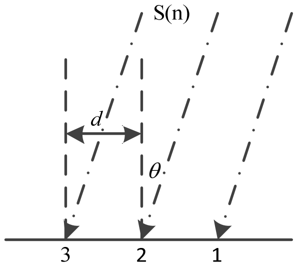

Fig.1. Schematic diagram of the principles of matched-signal phase method of three-sensor array.

Multiply Both sides of equation (18) and (19) by $e^{j\omega\tau}$ and $e^{j2\omega\tau}$:

(20) $$\left|X_1(j\omega)\right|e^{j\psi_1}=\left|S(j\omega)\right|e^{j\alpha}+\left|W_1(j\omega)\right|e^{j\varphi_1}$$

(21) $$\left|X_2(j\omega)\right|e^{j(\psi_2+\omega\tau)}=\left|S(j\omega)\right|e^{j\alpha}+\left|W_2(j\omega)\right|e^{j(\varphi_2+\omega\tau)}$$

(22) $$\left|X_3(j\omega)\right|e^{j(\psi_2+2\omega\tau)}=\left|S(j\omega)\right|e^{j\alpha}+\left|W_3(j\omega)\right|e^{j(\varphi_3+2\omega\tau)}$$

With the method reference [11] proposed to find out the solution for the desired signal:

(23) $$\text{Re}(S)=\frac{EA-FB}{2(CA-DB)}$$

(24) $\mathrm{Im}(S)=\dfrac{FC-ED}{2(CA-DB)}$

Where,

$A=\mathrm{Im}(X_3-X_1)$, $B=\mathrm{Im}(X_2-X_1)$, $C=\mathrm{Re}(X_2-X_1)$, $D=\mathrm{Re}(X_3-X_1)$, $E=|X_2|^2-|X_1|^2$, $F=|X_3|^2-|X_1|^2$ .. Re represents the real part and Im represents the imaginary part.

In this way, the signal X_1 , X_2 and X_3 can estimate the speech signal $s(n)$ more accurately. Such treatment could not only get TEO which is more representative of the speaker features but also improve accuracy of CFCC characteristic parameter in low SNR environments.

The principle diagram of TEO-CFCC characteristic parameter extraction method proposed by this paper is shown in figure 2:

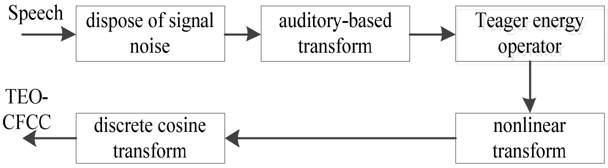

Fig.2. Schematic diagram of TEO-CFCC characteristic parameter extraction method

Firstly, dealt with the speech signal by using phase information to reduce the influence caused by noise signal. Secondly, dispose speech signal by using auditory-based transform, get power spectrum value $P(i)$. And then get calculation for TEO in each point of power spectrum according to the equation (14), get $y(i,j)$ through deal with the spectrum value after TEO transform by using nonlinear transform for equation (11)-(13). Finally, remove all of the correlation between signals by using the discrete cosine transform and get the speaker's TEO-CFCC characteristic parameter by mapping the signal to the low dimensional space.

3 The Experimental Results and Analysis

This paper use 90 speakers in the list of Train under the TIMIT speech database as test data, in which the number of female is 30 and the male is 60. This paper use Gaussian Mixture Model (GMM) which has nothing to do with this text as speech acoustic models. In order to get the effectiveness difference of the TEO-CFCC characteristic parameter, the pure environment, 5dB, 0dB, -5dB and -10dB noise environments are use to perform contrast test. The addition noise is the Vehicle interior noise, babble noise and white noise in the standard library noise (noisex-92). Figure 3-figure 5 shown the text results in the three noise conditions above.

As we can see from figure 3, three characteristic parameters achieve over 96% accuracy in clean testing conditions. But when white noise is gradually added to the clean testing data, the accuracy of MFCC characteristic parameter declines sharply. The accuracy of the MFCC is less than 50% when the SNR is 6dB, while accuracy of CFCC characteristic parameter and TEO-CFCC characteristic parameter proposed by this paper are close to 90%. When the SNR is 0 dB, the accuracy of the MFCC is less than 10%. Although the accuracy of CFCC characteristic parameter decline, it also reaches 60%. The accuracy of TEO-CFCC characteristic parameter proposed in this paper may reach 80% in 0 dB conditions. When the SNR is -5dB, the accuracy of the CFCC is less than 20%, while the TEO-CFCC characteristic parameter still has an accuracy of 57.8%.

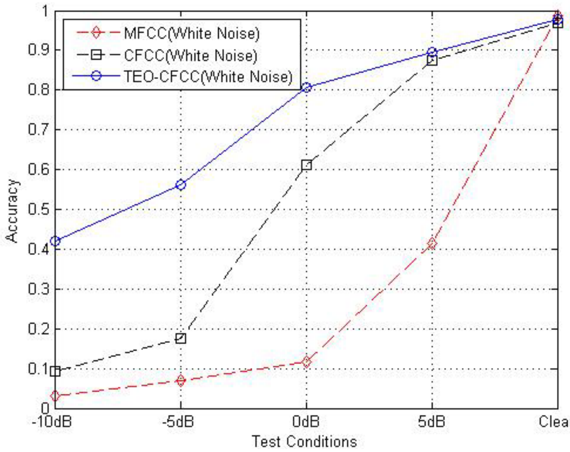

Fig. 3. Comparison diagram of the accuracy of MFCC, CFCC and TEO-CFCC characteristic parameter in white noise environment

As shown in figure 4 and figure 5, when the SNR is below 5dB, the accuracy of MFCC characteristic parameter decline sharply in the vehicle interior noise and babble noise testing conditions. Both CFCC characteristic parameter and TEO-CFCC characteristic parameter have a good performance of anti-noise, especially in the conditions in which the SNR is below -5dB. The accuracy of TEO-CFCC characteristic parameter is obviously better than CFCC characteristic parameter's. When the SNR of vehicle interior noise is -10dB, the TEO-CFCC characteristic parameter still has an accuracy of 70%.

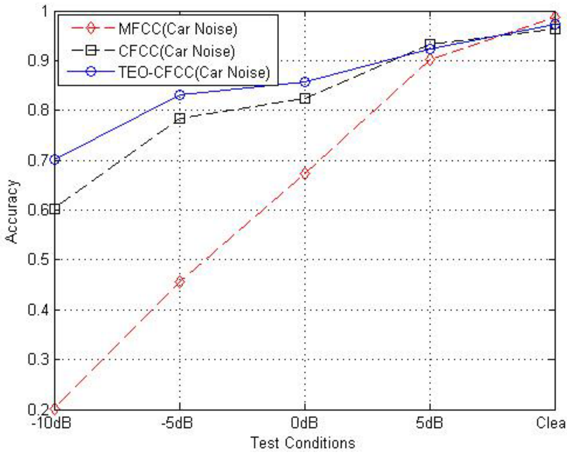

Fig. 4. Comparison diagram of the accuracy of MFCC, CFCC and TEO-CFCC characteristic parameter in vehicle interior noise environment

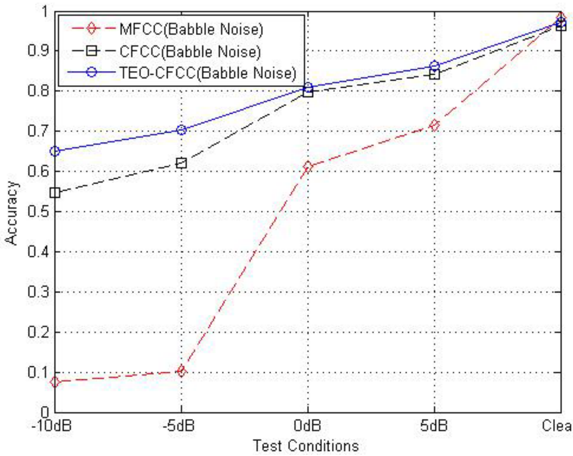

Fig. 5. Comparison diagram of the accuracy of MFCC, CFCC and TEO-CFCC characteristic parameter in babble noise environment

4 Conclusion

The phase matching method is combined with Teager energy operator on the basis of CFCC characteristic parameter is presented in this paper, and TEO-CFCC parameter method which can characterize personality characteristics is as well put forward. Tests which used the unified GMM speaker recognition model make it well known that the accuracy of TEO-CFCC characteristic parameter combined with the energy operator is obviously better than MFCC and CFCC. Experiment results show that the recognition accuracy can reach to 83.2% in a -5dB SNR of vehicle interior noise environment by using TEO-CFCC characteristic parameter.

REFERENCES

[1] Dulas J. Automatic word's identification algorithm used for digits classification [J]. *Przeglad Elektrotechniczny*,87(2011), No.11, 230-233.

[2] Dulas J. Speech signal's automatic segmentation based on the grid with various parameter's method [J]. *Przeglad Elektrotechniczny*,86(2010), No.1, 229-232.

[3] Dobrowolski A P,Majda E. Evaluation of the usefulness of selected features of the speech signal for automatic speaker recognition systems [J]. *Przeglad Elektrotechniczny*, 87(2011) No.10, 193-197.

[4] Vijayasenan D,Valente F,Bourlard H. Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features. *Speech Communication*, 54(2012) , No.7, 55-67.

[5] Wang L,Minami K,Yamamoto K,et al. Speaker Recognition by Combining MFCC and Phase Information in Noisy Conditions. *IEICE Transactions on Information and Systems*, E93D(2011), No.9, 2397-2406.

[6] Li Q,Huang Y. An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions. *IEEE Transactions on Audio Speech and Language Processing*, 19(2011), No.1, 1791-1801.

[7] Qi L. An auditory-based transfrom for audio signal processin. *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY, United states.Oct18-21, (2009). 181-184.

[8] Dimitriadis D,Maragos P,Potamianos A. On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition. *IEEE Transactions on Audio Speech and Language Processing*, 19(2011), No.4, 1504-1516.

[9] Tu C-C,Juang C-F. Recurrent type-2 fuzzy neural network using Haar wavelet energy and entropy features for speech detection in noisy environments. *Expert Systems With Applications*, 39(2012), No.7, 2479-2488.

[10] Lou Hongwei,Hu Guangrui. Speech Feature Based on Teager Energy Operator and Dyadic Wavelet Transform, *Journal of Shanghai Jiaotong University,* 37(2009), No.2, 83-85.

[11] Sun Jincai, Zhu Weijie,Sun Tieyuan. DOA and Waveform Estimation by Using Small-Dimension Array, *Journal of Northwestern Rolytechnical University*, 21(2003), No.4,512-515.

***Authors**: Jingjiao Li. School of Information Science & Engineering, Northeastern University, Shenyang, China*

Dong An. School of Information Science & Engineering, Northeastern University, Shenyang, China, E-Mail: 249350656@qq.com

Dan Zhao. Department of Neurosurgery, The First Affiliated Hospital of China Medical University,Shenyang, China

Caoqun Rong, School of Information Science & Engineering, Northeastern University, Shenyang, China

Shuang Ma, School of Information Science & Engineering, Northeastern University,Shenyang, China