Send Orders for Reprints to reprints@benthamscience.ae

The Open Automation and Control Systems Journal, 2014, *6*, 1661-1668

The Self-adaptive Voice Activity Detection Algorithm based on time-frequency Parameters

Xiaohua Wang* and Lei Qu

College of Electronics and Information, Xi 'an Polytechnic University, Xi 'an 710048 China

Abstract: In order to solve the inferior performance and sad self-adaptive of the traditional voice activity detection algorithm in an environment with low Signal to Noise Ratio (SNR), a new self-adaptive voice activity detection algorithm based on time-frequency (TF) parameters is put forward. After introducing the time-domain log-energy and improved mel-scale log-energy, the new TF parameters are acquired by coalescing them, which make it possible for distinguishing speech from noise effectively. Then, the TF parameters are updated to predicate endpoint through the threshold test. Finally, simulation experiments show that the algorithm can improve significantly the performance of automatic speech recognition (ASR) system and robustness. When the SNR is 0dB, the error rate of the algorithm is about 15%.

Keywords: Self-adaptive, voice activity detection, mel-scale log-energy, TF parameter.

1. INTRODUCTION

In the last few years, the research of Automatic Speaker Recognition (ASR) has made a rapid development [1, 2], particularly in Voice Activity Detection (VAD), which is the basic part of ASR [3]. Accurate VAD is able to distinguish speech from noise better to simplify successor activities in ASR system such as feature extraction and recognition, not only improving the running speed but also the recognition performance of the overall system effectively [3, 4]. Therefore, the VAD is crucial in ASR system.

The current VAD can be roughly divided into several categories. (1) VAD based on time-domain parameters [5, 7] such as short-time energy, zero-crossing rate, log-energy and auto-correlation function, they are not calculated in an environment with low SNR because of the bad robustness and noise immunity in spite of the simple computations; (2) VAD based on frequency-domain parameters covering wavelet entropy [8], Mel cepstrum distance [9] and sparse power spectrum [10], they own sad accurate rates of detection in an environment with low SNR albeit the better immunity and simple implementation; (3) VAD based on models embracing support vector machine [11] and neural network [12], they are analogical with model matching in the last part of ASR system, that is modeling respectively for speech and noise, and they own larger complexity as well as the noise model built may not compatible with the realistic environment on account of the fickle noise; (4) VAD based on a variety of new theories incorporating chaos theory[13] and fractal theory[14], they are merely appropriate for some individual noises by reason of the complexity and limitations.

Nonetheless VAD algorithm based on fusion of short-time energy and zero-crossing rate has been proposed by Sun Zhanxian [7], the algorithm has an defect, that is the poor noise immunity of the time-domain parameters. Thereby a new VAD algorithm based on time-frequency (TF) parameters is inspired by coalescing the log-energy and frequency-domain parameters. When extracting the frequency-domain parameters, the multiband analysis of mel-scale frequency bank and order statistic filter (OSF) have been merged to acquire mel-scale log-energy [9, 15]. Then the TF parameters obtained contain a good deal of information to predicate endpoint by dynamic threshold. In this way, the endpoint determination is precise. Finally, the algorithm presented in this paper and in literature [7] are added to the ASR system respectively, then make comparation with the system without any VAD. It can be easily found that the proposed algorithm significantly improves the recognition rate of the ASR system. The experimental results show that the algorithm proposed in this paper improves the exactitude and effectiveness in an environment with low SNR. Compared with the VAD algorithm based on time-domain fusion[7] and VAD algorithm based on frequency-domain parameters[9], the algorithm proposed in this paper performs better.

2. BASIC IDEA OF THE ALGORITHM

The nature of VAD is to distinguish different characteristics of voice segment and non-voice segment in the time domain, frequency domain or transform domain, including preprocessing, feature extraction, feature confusion and endpoint adjudication. This paper focuses on improving the last three aspects. Here is the basic idea of the algorithm idea.

(1) In feature selection, it's better to choose two characteristic parameters whose advantages and disadvantages are complementary;

1874-4443/14

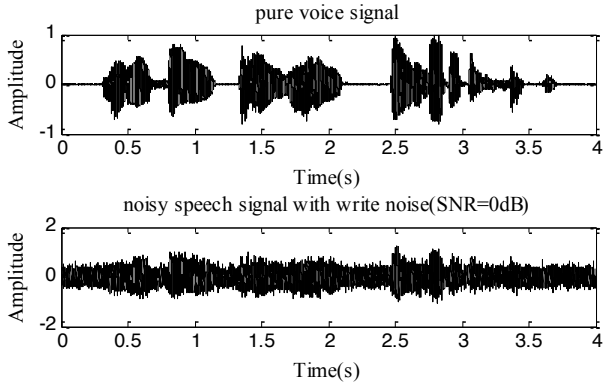

Fig. (1). Pure Voice Signals and Noisy Speech Signal Curve.

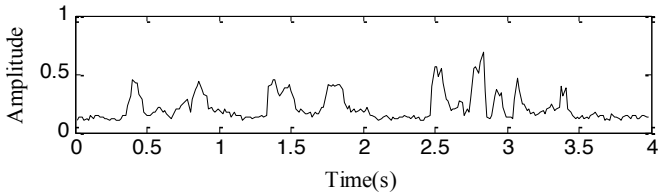

Fig. (2). Characteristic Curve from Log-energy.

(2) In feature extraction, it should be possible to reduce the effects of noise and improve the adaptive ability of the algorithm;

(3) In feature confusion, the advantages of the two features should be highlighted, avoiding the mistakes caused by feature that is not obvious enough.

(4) In endpoint adjudication, dynamic threshold is used to improve the accuracy and adaptability of the algorithm in different noise.

This article improves the traditional endpoint detection algorithm based on the principles above.The feature extraction in tradition is to fetch single parameters such as time-domain parameters or frequency-domain parameters as features to distinguish speech from noise. However, in this paper, time-domain parameters and frequency-domain parameters are merged to conquer the shortcoming that features from single parameters own inferior noise immunity and discrimination. Otherwise, the endpoint determination in tradition uses a fixed threshold, while a dynamic threshold in accordance with the noise characteristics in this paper is calculated to surmount that the fixed threshold adapts environment weakly.

3. IMPROVED VAD ALGORITHM

3.1. Time-domain Log-energy

In endpoint detection, short-time energy, zero-crossing rate, short-time average amplitude and auto-correlation function are common features in time domain parameters. A log-energy feature has been proposed [6], and compared with short-time energy, the log-energy makes it possible for better distinguishing speech from noise in place of confusing mute and voiceless with small amplitude and too large noise characteristics. Besides, the log-energy feature extraction is relatively simple. Therefore, it's quite preponderant to ultimate the log-energy feature in this paper. Here is the calculation.

Suppose time-domain signal of the speech is $x(i)$, then the speech signal $x_n(i)$ in the n'th frame is calculated after preprocessing by window function $w(n)$.

$$x_n(i) = x\left((n-1)\times inc + L\right)w(n), 1 \le i \le L, 1 \le n \le fn \tag{1}$$

Where $w(n)$ is Hamming window; inc($inc < L$) is the frame shift, L is frame length, fn is the total number of frame after framing.

In order to keep the continuity of frames, it's necessary to retain some overlaps between the continue frames.

The equation of log-energy of $x_n(i)$ is shown as follows.

$$\begin{aligned} LE(n) &= \log_{10}\left(E(n)+a\right) - \log_{10}(a) \\ E(n) &= \sum_{i=1}^{L} x_n^2(i) \end{aligned} \tag{2}$$

Where $E(n)$ are the nth short-time energy; a is a constant. Suppose $a = 1$.

Suppose a write noise is mixed into a 4s pure voice signal as a noisy speech signal with 0dB SNR by equation (3), and Fig. (**2**) shows the curves of pure voice signal and noisy speech signal.

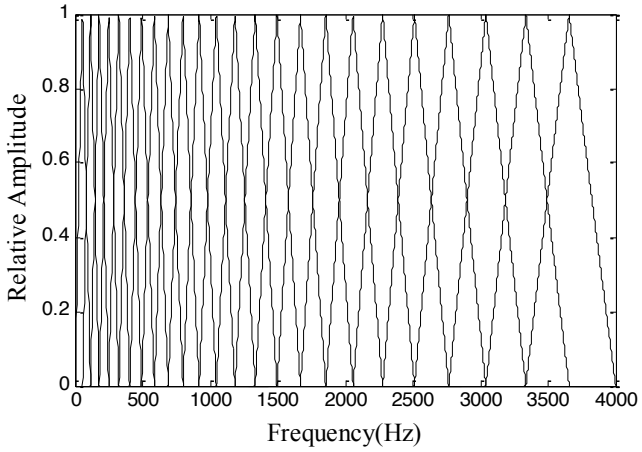

Fig. (3). Frequency-response Curve from Mel-scale Filter Bank.

$$SNR = 10\log_{10}\frac{power(signal)}{power(k \bullet noise)} \qquad (3)$$
$$X = signal + k \bullet noise$$

Where $power(signal)$ is voice signal energy; $power(k \bullet noise)$ is nose signal energy; k is a proportional coefficient; X is the noisy speech signal.

Then the time-domain log-energy are extracted according to equation (2) and the characteristic curve from log-energy is shown in Fig. (**2**).

From Fig. (**2**), it can be easily find that it's difficult to differentiate unvoiced part from noisy part of time-domain log-energy in an environment with low SNR, While the dullness parts are relatively clear.

3.2. Mel-scale Log-energy

A triangular filter bank called mel-scale filter bank is required to imitate the nonlinear frequency domain that human perceive, and the filter bank is uniform distribution in mel-scale frequency. The relationship between mel-scale frequency and frequency (Hz) is described as follows.

$$F_{mel} = 2595\ln(1 + f/700) \qquad (4)$$

Where F_{mel} is perceived frequency with *mel* unit; f is the actual frequency with *Hz* unit.

Feature extraction of mel-scale frequency energy is divided into the following steps:

(1) Pre-emphasis

The purpose is to compensate for the loss of the high frequency portion, so that the signal becomes flat for facilitating frequency analysis. It is generally an order FIR filter:

$$H(z) = 1 - \mu z^{-1} \qquad (5)$$

Where, μ ($0.9 < \mu < 1.0$) is the pre-emphasis coefficient, and suppose μ =0.9375 in this paper.

Suppose $x(i)$ is the value of voice sampling in the i'th time, then the result is $y(i) = x(i) - \mu x(i-1)$.

(2) Short-time Fourier transform (STFT)

Since the speech signal is stationary shortly, it's needed framing process of the speech signal and calculation of Fourier transform in each frame. In this way, the STFT is obtained.

$$X_n(k) = \sum_{i=0}^{N-1} y(i) w(n-i) e^{-j\frac{2\pi}{N}ki}, 0 \le k \le N-1 \qquad (6)$$

Where, $y(i)$ is a speech sequence after the pre-emphasis.

For frequency-domain analysis of the speech signal, the shape of the window function is very important. The spectral of rectangular window has better smoother performance, but some details of its waveform are lost so that more serious leakage phenomenon occurs. On the other hand, the Hamming window is able to overcome the phenomenon effectively, which is the reason to choose Hamming window.

$$w(n) = \begin{cases} 0.54 - 0.64\cos\left[2\pi n/(N-1)\right], & 0 \le n \le N-1 \\ 0 \end{cases} \qquad (7)$$

Where, N is the window length.

(3) Spectral energy

After STFT, spectral energy in each frame is calculated as follows.

$$E_n(k) = \left|X_n(k)\right|^2 \qquad (8)$$

(4) Mel-scale sub-band energy

Fig. (**3**) is frequency-response curve from mel-scale filter bank designed in the light of mel-scalefrequency. In the range of $0 \sim 4000$Hz , the filter bank embraces 24 triangular

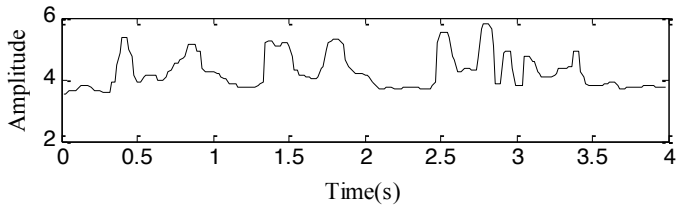

Fig. (4). Characteristic Curve from Mel-scale Log-energy.

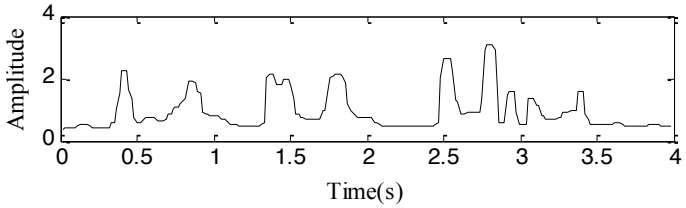

Fig. (5). Characteristic Curve from TF Parameters.

band-pass filers, $f(m,k)$, where $1 \le m \le 24$, $0 \le k \le 256$, then each filter has a triangular band-pass frequency response, and the granularity of bandwidth is confirmed by the mel-scale frequency interval from equation (4). The values of $f(m,k)$ denote the weighting factor of the frequency energy at the *k*th point of the *m*th sub-band. So the mel-scale sub-band energy at the the *n*th frame of the *m*th sub-band:

$$S(n,m) = \sum_{k=0}^{N-1} E_n(k) f(m,k) \tag{9}$$

Before calculating the current mel-scale log-energy, the mel-scale sub-band energy is needed to smooth, owing to $S(n,m)$ have a large fluctuation in an environment with low SNR and affect the accuracy of VAD. The process of smooth adopts OSF in this paper.

(5) Energy after OSF filtering

OSF is firstly used in image edge detection, and Ramirez is the first person applying OSF to VAD [15]. It's good news that OSF has the ability of improving accuracy.

Mel-scale sub-band energy { $S(n-N,m)$, … , $S(n,m)$, … , $S(n+N,m)$ }with the length of $L(L=2N+1)$ are ascending sorted as $S_{(h)}(n,m)$ $(1 \le h \le L)$, where n is the number of the current voice frame, the beginning of voice signal with length of N ($N=5$) frames is supposed to be pure noise. Then mel-scale sub-band energy will be filtered by OSF to receive mel-scale sub-band energy $S_p(n,m)$ at the *m*th sub-band of the *n*th frame as

$$\begin{aligned} S_p(n,m) &= (1-f) S_{(h)}(n,m) + f S_{(h+1)}(n,m) \\ h &= \lfloor 2pN \rfloor \text{ } f = 2pN - h \end{aligned} \tag{10}$$

Where $\lfloor \ \rfloor$ is to rounding the decimal portion; p is sampling quantile of *OSF* with Gauss distribution. Suppose $p = 0.9$.

Finally, the nth mel-scale log-energy is give by

$$MLE(n) = \log \sum_{m=0}^{M-1} S_p(n,m) \tag{11}$$

The mel-scale log-energy are extracted from a noisy speech signal with write noise and 0dB SNR, then the characteristic curve is shown in Fig. (**4**).

As can be seen from Fig. (**4**), in an environment with low SNR, mel-scale log-energy is better to distinguish unvoiced part from noisy part and the slope of the curve is large enough to shun false judgment when the transition from noisy segment to speech segment.

3.4. TF Parameters

The obtained time-domain log-energy and mel-scale log-energy are integrated as a new TF parameters

$$TF(n) = smooth\left(c \bullet LE(n) \times MLE(n)\right) \tag{12}$$

Where *smooth*( ) is performed by a three-point mean filter; constant c is a proper weighting factor. Suppose $c = 1$.

The TF parameters are extracted from a noisy speech signal with write noise and 0dB SNR, then the characteristic curve is shown in Fig. (**5**).

As can be seen from Fig. (**5**), in an environment with low SNR, The TF parameters not only assimilate the merits of time-domain log-energy and mel-scale log-energy, but also possess a flat curve in the noisy segment, which make it possible for selecting threshold easily and enhancing the accuracy of endpoint determination.

3.5. Dynamic Threshold

A new TF parameter will be created in each frame signal after fusion. Then the beginning of voice signal with length of N frames are adhibited to initialize threshold as

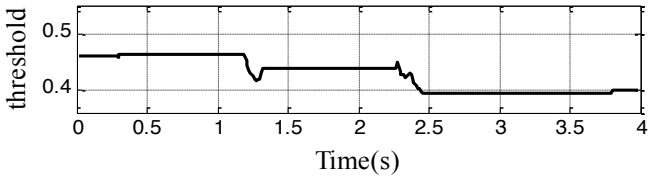

Fig. (6). Curve of Dynamic Threshold.

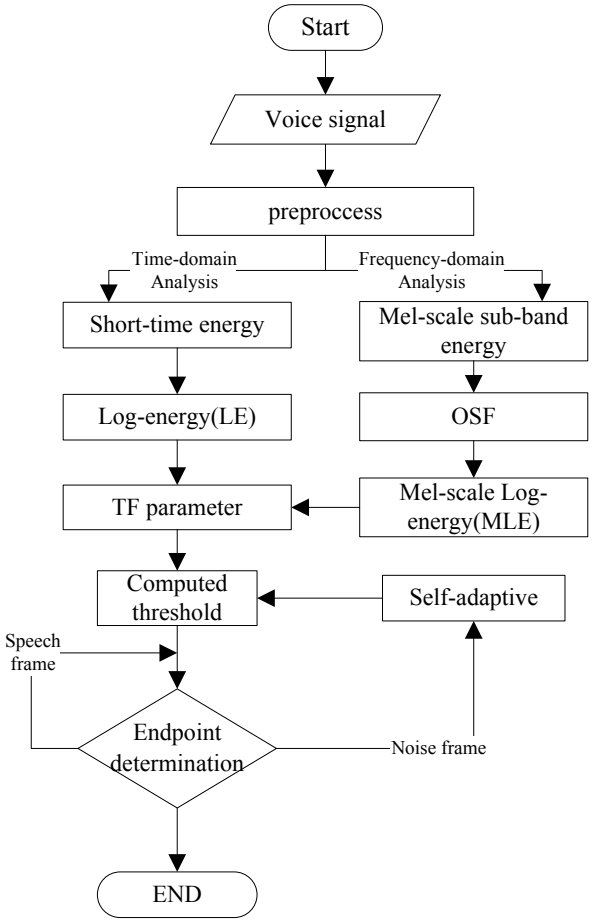

Fig. (7). The Whole Process of The Algorithm.

$$T = \alpha TF_N + \beta \tag{13}$$

Where α and β denote constant coefficients selected by modulation according to experiment results or curve fitting, suppose $\alpha = 1.25$ and $\beta = 0.01$; $TF_N = mean(TF(N))$, where $mean()$ denotes arithmetic average; T needs dynamic update for accurate endpoint determination.

Each TF parameter needs threshold detection: if the parameter is greater than T, the signal is determined as speech frame; Or noisy frame. In this way, TF_N can be updated as TF_{new}

$$TF_{new} = \left(TF_N \times l + TF\left(n\right)\right) \Big/ \left(l+1\right) \tag{14}$$

Where l is a constant. Suppose l=9.

Then updated TF_{new} is substituted into equation (15) to update threshold value.

$$T = \alpha TF_{new} + \beta \tag{15}$$

Finally, the dynamic threshold can be worked out. The curve of dynamic threshold in a noisy speech signal with factory noise and 0dB SNR is shown in Fig. (**6**).

The whole process of the algorithm proposed in this paper is shown in Fig. (**7**).

4. EXPERIMENTS AND DISCUSSION

The VAD algorithm proposed in this paper is simulated through Matlab.

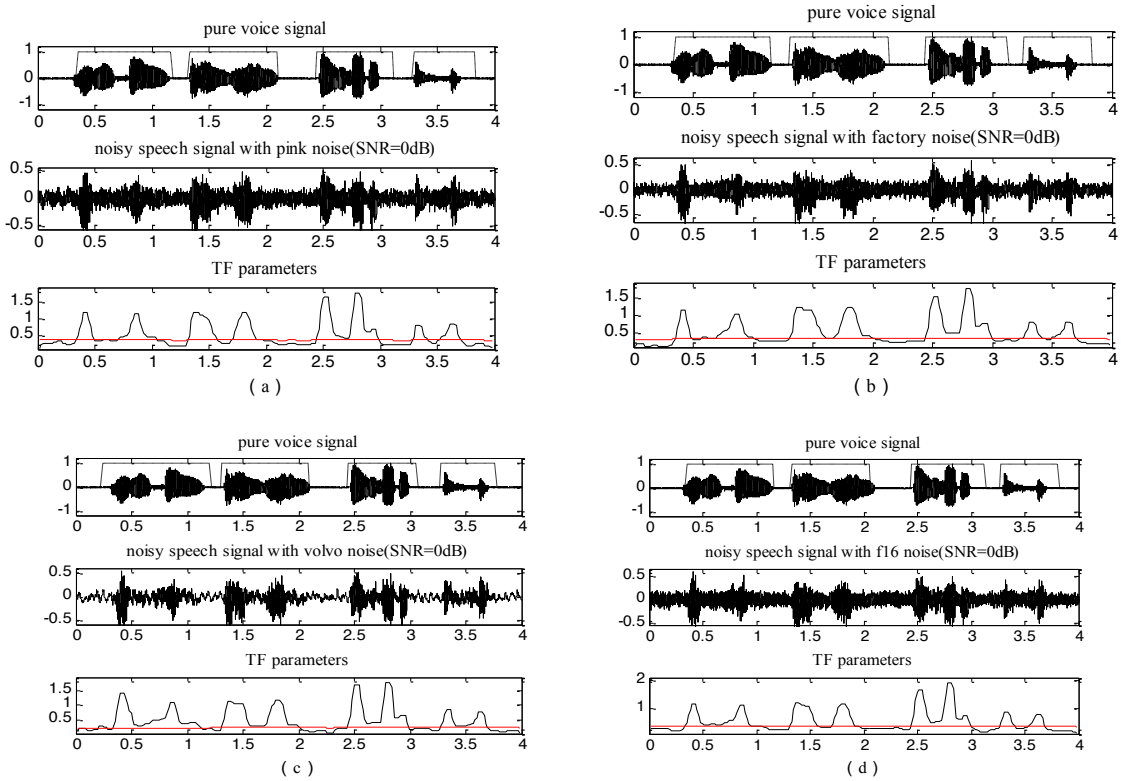

Fig. (8). Endpoint Detection Result.

4.1. The Influence of The Algorithm to ASR System

4.1.1. The Experimental Configuration

In the experiment, the non-speech frames that detected from the algorithm proposed in this paper are discarded directly, instead of sending to the back-end recognizer. In the whole experiment, the parameters in the recognizer are set constantly. The extracted features are 39-dimension MFCC, which consist of MFCC with 13-dimension, the first-order difference with 13-dimension, the second-order difference with 13-dimension. Let's adapt DTW to the Speaker Recognition, which contains specific phrases.

Then some voice sample should be collected as training templates and testing templates. For example, 40 people pronounce some words 10 times and these voices are recorded as training templates. Then 40 groups of training templates come out. On the other hand, these people pronounce some other words 5 times again and these voices are recorded as testing templates. Then 40 groups of testing templates come out.

4.1.2. Experiments and Discussion

In order to prove the algorithm presented in this paper, what can improve the accuracy of ARS. Three recognition rates are shown in Table **1**, which are calculated respectively from ASR system without any VAD algorithm (MFCC-DTW), ASR system with VAD algorithm based on fusion of short-time energy and zero-crossing rate (SZ-MFCC-DTW), ASR system with the algorithm proposed in this paper (TF-MFCC -DTW).

From Table **1**, it's obviously found that the recognition rate of ASR system is clearly increased after using the VAD algorithm proposed in this paper. The interferential frame are reduced after using VAD algorithm, which makes the training templates more effective and the recognition results more accurate.

4.2 Accuracy of The Algorithm

4.2.1. The Experimental Configuration

In a quiet environment, voice of 10 people including 5 men and 5 women with length of 4s is recorded as voice signals. The sampling rate is 8KHz, quantification is 16bit, and the noisy signal is from the NOISEX-92 database. Adobe Audition is utilized as the tooldif of speech signal analysis, and the pure speech signal endpoints are labeled manually as the standard detection.

4.1.2. Experiments and Discussion

To evaluate the robustness of the algorithms, the voice signals are respectively mixed with *pink* noise, *factory* noise, *volvo* noise and *f16* noise as noisy speech signals, which cover the *SNR* as *-5dB, 0dB, 5dB, 10dB* and *20dB*. Then compare each other. Considering the similarity of the comparisons, the endpoint detection results of the noisy speech signals with *0dB SNR* are enumerated in Fig. (**8**).

It can be drawn from the results, in a noisy environment *with 0dB SNR*, the proposed algorithm is still valid. For the preprocessing of ASR system, the enthesis of voice signal is

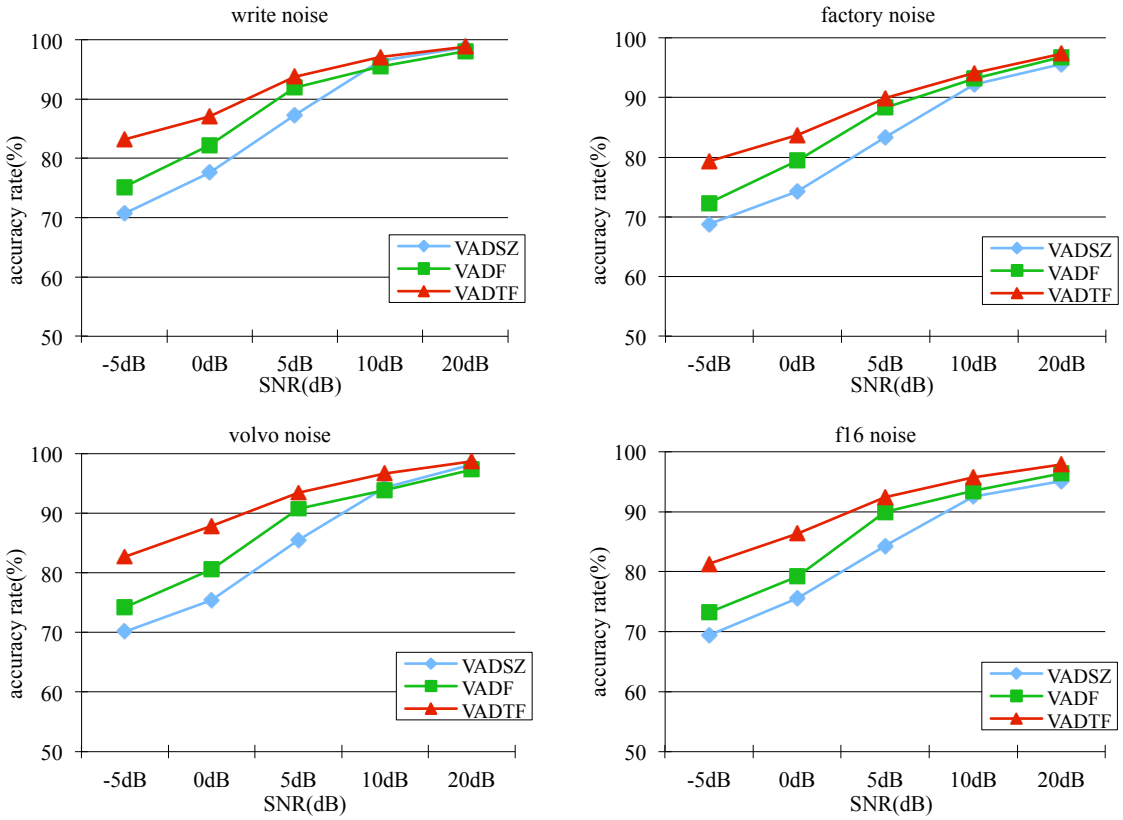

Fig. (9). Results of Each Algorithm in Different SNR.

needed to detect, instead of words segmentation, thereby some intermediate detection errors such as false detection and missed detection are not concerned. So the detection results in Fig. (**8**) are perfectly acceptable, which further demonstrate the algorithm works effectively and owns nice robustness in the environment with multiple noise.

Table 1. Recognition rates form 3 kinds of ASR system.

ASR System	Recognition Rate
MFCC-DTW	89.7%
SZ-MFCC-DTW	93.7%
TF-MFCC-DTW	95.3%

For the sake of testing the accuracy of this algorithm, a large number of experiments have been conducted by VAD algorithm based on fusion of short-time energy and zero-crossing rate [7] (we call it VADSZ), VAD algorithm based on frequency parameters [9] (we call it VADF) and the algorithm proposed in this paper(we call it VADTF). Then the reference of the detection accuracy rate is R by equation (17). The results of each algorithm are presented in Fig. (**9**).

$$R = \frac{sum\ of\ the\ number\ of\ \text{correctly}\ detected\ voice}{sum\ of\ the\ number\ of\ voice\ labelled\ manually} \quad (16)$$

A conclusion can be put forward through the comparative analysis in Fig. (**9**) is that the VAD algorithm proposed in this paper has higher accuracy rate under conditions of different noise and different *SNR*, especially under conditions of low *SNR*. On the other hand, the accuracy under smooth noise such as write noise is just about 2% higher than other non-stationary noise, which indicates that the dynamic threshold for endpoint determination is appropriate for dealing with non-stationary noise and valuable in application.

CONCLUSION

The algorithm proposed in this paper has improved traditional algorithm. It finally calculated new parameters via the fusion of time-domain log-energy and mel-scale log-energy. These parameters made the features of voice segments and noise segments obviously and selected threshold briefly. Otherwise, dynamic threshold in the algorithm has advanced the adaptive ability and accuracy rate of endpoint determination. On the other hand, mel-scale sub-band energy is a portion of feature selection in ASR system, hence the algorithm blend in the ASR system easily improve the performance of speaker recognition systems.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors would like to thank The National Natural Science Foundation of China (Program No. 61301276); Startup Project of Doctor scientific by the Xi'an Polytechnic University (BS1207); Xi 'an Polytechnic University Discipline Construction Funded Projects (107090811).

REFERENCES

[1] P. Li and H. Tang, "Design of a Low-Power Coprocessor for Mid-Size Vocabulary Speech Recognition Systems", *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 58, pp. 961-970, May 2011.

[2] Z. Li, W. Zhang and H. Liang, J. Liu, "Total Variability Subspace Adaptation Based Speaker Recognition", *ACTA AUTOMATICA SINICA,* vol. 40, pp. 1836-1840, Aug 2014.

[3] L. Guo, X. He Y. Zhang and Y. Lv, "An Order Statistics Filtering-based Real-time Voice Activity Detection Algorithm," *ACTA AUTOMATICA SINICA*, vol. 34, pp. 419-425, Apr 2008.

[4] J.C. Junqua, B. Mak and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", *Speech and Audio Processing, IEEE Transactions on*, vol. 2,pp. 406-412, Jul 1994.

[5] X. Yang, B. Tan, J. Ding, J. Zhang and J. Gon, "Comparative Study on Voice Activity Detection Algorithm", *in Electrical and Control Engineering (ICECE), 2010 International Conference on*, 2010, pp. 25-27.

[6] H. zhao, G. Wang and L. zhao, "A New Voice Activity Detection Using Logarithmic Energy Spectral Entropy", *Journal of Hunan University: Natural Sciences*, vol. 3, pp. 72-77, Jul 2010.

[7] Z. Sun, F. Huang and J. Wang, "Self-adaptive Algorithm of Voice Endpoint Detection", *Computer Engineering and Applications,* vol. 50, pp. 206-210, Jan 2014.

[8] G. Lee, S. D. Na, J. H. Cho and M. N. Kim, "Voice Activity Detection Algorithm Using Perceptual Wavelet Entropy Neighbor Slope", *Bio-medical materials and engineering,* vol. 24, pp. 3295-3301, Jun 2014.

[9] Z. Chen, W. Wu, J. Liu and S. Xia, "Voice Activity Detection Algorithm based on Mel Cepstrum Distance Order Statistics Filter", *Journal of University of Chinese Academy of Sciences,* vol. 31, pp. 524-529, Jul 2014.

[10] D. You, J. Han, G. Zheng and T. Zheng, "Sparse Power Spectrum based Robust Voice Activity Detector", *in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on,* 2012, pp. 25-30.

[11] J. Wu and X. Zhang, "Efficient Multiple Kernel Support Vector Machine based Voice Activity Detection", *Signal Processing Letters, IEEE,* vol. 18, pp. 466-469, Aug 2011.

[12] X. Zhang, J. Wu, "Deep Belief Networks based Voice Activity Detection", *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 697-710, Apr 2013.

[13] T. Song, K. Lee and H. Ko, "Robust Visual Voice Activity Detection Using Chaos Theory under Illumination Varying Environment", *Consumer Electronics (ICCE), 2014 IEEE International Conference on,* 2014, pp. 10-13.

[14] S. Shafiee, F. Almasganj, B. Vazirnezhad and A. Jafari, "A Two-Stage Speech Activity Detection System Considering Fractal Aspects of Prosody", *Pattern Recognition Letters,* vol. 31, pp. 936-948, Jul 2010.

[15] J. Ramírez, J. C. Segura and C. Benítez, "An Effective Subband OSF-Based VAD With Noise Reduction for Robust Speech Recognition", *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 1119-1129, Nov 2005.

[16] Z. Song, Application of MATLAB in Speech Signal Analysis And Synthesis, Beijing, Beihang University Press, 2013.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014