AN ADAPTIVE VOICE ACTIVITY DETECTION ALGORITHM

Zhang Zhigang[1] and Huang Junqin[2]

[1]School of Printing and Packaging Engineering, Xi'an University of Technology, Xian, China

[2] Engineering Training Center, Xi'an University of Technology, Xian, China

Emails: [1]zzg@xaut.edu.cn, [2]HUANGJQ@xaut.edu.cn

***Submitted:* May.10, 2015** **Accepted: Nov. 10, 2015** **Published: Dec. 1, 2015**

Abstract- Voice Activity Detection (VAD) is a crucial step for speech processing, which detecting accuracy and speed directly affects the effect of subsequent processing. Some voice processing system based phone or in the indoor environment, which need simple and quick method of VAD, for these representative voice signal, this paper proposes a new algorithm which is adaptive and quick based on a major improvement to Dual-Threshold endpoint detection algorithm. First the amplitude normalization is processed to the original voice signal, the characteristic is extracted by means of short-time amplitude, which can simplify operation. Then, large-scale (long frame-length and frame-shift) short-time amplitude is used for rough detection, combining adaptive threshold judgement of consecutive frames, which can find voice areas of start-point and end-point quickly. To these areas, small-scale (short frame-length and frame-shift) short-time amplitude is used for accurate detection, forward scanning is put to start-point area, reverse scanning is put to end-point area, combining adaptive threshold judgement of consecutive frames, start-point and end-point of the effective speech can be accurately located. Experimental results show that the method of this paper can detect endpoints of voice signal more quickly and accurately, which can improve recognition performance dramatically. Large-scale can increase detection speed, small-scale can improve detection accuracy, both can be adjusted to satisfy the different requirements. The method of this paper ensures both detection speed and precision, which has more flexibility and applicability.

Index terms*:* Voice signal, Endpoint detection, Short-time amplitude, Multi-scale detection, Adaptive threshold.

I. INTRODUCTION

The purpose of VAD (Voice Activity Detection) is to detect the useful voice segment from the original audio signal, and to locate the starting point and ending point, which is very important for speech recognition, the researches show that more than 50% error of speech recognition originated from inaccurate VAD[1], thus it can be seen that rapid and accurate VAD is a vital process.

Rapid and accurate VAD can play a very important role in speech recognition, for example, it can eliminate the noise signal segment, and can extract feature to the speech signal segments merely, which not only reduces the amount of calculation to speed up the processing, but also can improve the recognition accuracy rate. In speech coding, it also can reduce the bit rate of noise and improve the coding efficiency without affecting the quality of speech signal.

In recent years, many methods of VAD have been put forward, which can be divided into many types.

Traditional VAD methods generally utilize cognitive or statistical characteristics of voice signal to distinguish speech and noise, which generally can be divided into three categories in accordance with feature extraction method: (1)Time-Domain detection method, such as short-time energy and zero-crossing rate of Dual-Threshold detection[2], Short-time Autocorrelation[3]; (2) Frequency-Domain detection method, such as LPC Cepstrum[4], Spectrum Entropy[5], Energy spectrum entropy[6-7], Distance Entropy [8], Sub-band Energy[9], Harmonic Energy[10], etc. (3)Nonlinear feature detection method, such as Permutation Entropy[11], C0 measure[12], HHT[13] , etc. Traditional VAD methods have been widely researched and applied because of their simple, but their limitations are also obvious, which detection effect easily affected by noise, they can’t adapt to different environments especially the low-SNR condition.

Then, the later scholars utilize the combination of a variety of characteristics and adopt pattern recognition methods for VAD, many new algorithms appear constantly, such as [14] proposed a VAD approach Mel Frequency Cepstrum Coefficient (F-MFCC) based on Fisher linear discriminant analysis, it approach achieves higher VAD accuracy under different SNR and noise conditions.[15] proposed to fuse multiple features via a deep model, called deep belief network (DBN), [16] presents an algorithm based on SVM and Wavelet Analysis, [17] adopted neural networks, [18] proposed a method in a kernel subspace domain to improve the performance of

the kernel-based VAD, [19] used real-valued neural network (RVNN) to estimate the coefficients of an autoregressive (AR) model, [20] proposed a statistical voice activity detection method in a high-dimensional kernel feature space by a nonlinear mapping, [21] eliminated the influence of noise by means of entropy-based measure. [22] proposed a novel feature extraction algorithm based on the double-combined Fourier transform and envelope line fitting is proposed. [23] extracted feature using an equivalent rectangular bandwidth (ERB) filter band cepstrum and constructed a learning model using the acoustic model to improve the speech detection and recognition, etc. Those methods take the advantages of various characteristic and intelligence algorithm, improve the detection effect under noise environment effectively, but their algorithmically complex increased sharply, and the detection effect remains unsatisfactory under low SNR, especially for non-stationary noise.

Recent study on VAD is more closer to specific application background, such as [24] proposed a framework which attempted to incorporate articulatory information into the stochastic segment model based on Mandarin speech detection and recognition system,[25] researched on Hindi speech, adopts Mel frequency cepstral coefficients(MFCC), Perceptual linear prediction coefficients (PLP) and PLP derived from Mel-scale filter bank (MFPLP), combines Auto-associative neural networks (AANN) to detect and recognize Hindi speech.

But it now appears there isn't a perfect method of VAD which can separate accurately speech and noise under low SNR, it needs our unceasingly research and the exploration.

We should choose the appropriate method of VAD according to the actual situation, for instance, in real life the voice signal is collected by microphone in many situations, which SNR usually is acceptable, therefore we can adopt quick and easy method. In that case, simple and fast method is a good choice. The typical VAD method is Dual-Threshold detection method, which has high detecting speed because of its simple, for these reasons, this paper made some improvement on it and proposes a new algorithm, which use short-time amplitude as characteristics, large-scale(long frame-length and frame-shift) was first used for rough detection, which can find quickly voice areas of start-point and end-point, then small-scale(short frame-length and frame-shift) is used for detecting the accurate location of start-point and end-point, which can scan those endpoint areas forward or reverse according to the type of endpoint, the endpoint of effective speech can be accurately located combined with adaptive threshold.

II. PREPROCESSING OF SPEECH SIGNAL

The speech segment used in this paper, part samples come from YOHO speech database, another were recorded by us in common indoor environment.

Speech signal is a typical time variant and non-stationary signal, but its characteristic keeps mostly unchangeable in a short time range (10~30ms), thereby short-time energy, short-time average amplitude, short-time zero crossing rate characteristics of the speech signal can be used for detection of speech signal.

The speech signal collected from indoor environment is most representative for speech recognition, which SNR is usually high. In this case, the energy of speech segment is usually greater than the energy of noise or silence segment obviously, which can used for endpoint detection.

Dual-Threshold endpoint detection algorithm often used for processing indoor speech signal because of its good detection results, it takes short-time energy and short-time average zero-crossing rate as characteristics to detect the endpoint of speech signal.

The formula of short-time energy is as follows:

$$E(n) = \sum_{i=1}^{N} x_n^2(i) \qquad (1)$$

In the equation, N is frame length, x(i) is original speech signal, E(n) is the short-time energy of frame n.

Short-time average zero-crossing rate means the number of alternative times of sampling point's amplitude value between positive or negative per unit time, the formula is as follows:

$$Z_n = \frac{1}{2}\sum_{m=1}^{N-1}\left|\mathrm{sgn}[x_n(m)] - \mathrm{sgn}[x_n(m-1)]\right| \qquad (2)$$

In above equation, sgn[] is sign function, which definition is as follows:

$$\mathrm{sgn}[x] = \begin{cases} 1, (x \geq 0) \\ -1, (x < 0) \end{cases} \qquad (3)$$

Dual-Threshold endpoint detection algorithm has two detection processes, first, a higher threshold (E_H) is set on the basis of short-time energy E(n), which can ensure most speech signal to pass the threshold, a probable range of speech signal can be designated, shown as "A-B segment" in Figure 1.

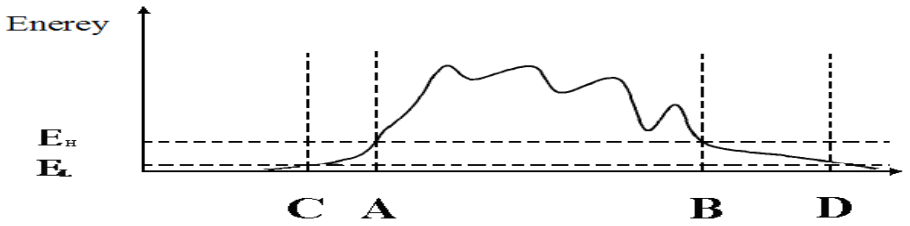

Figure 1. First process of Dual-Threshold endpoint detection algorithm

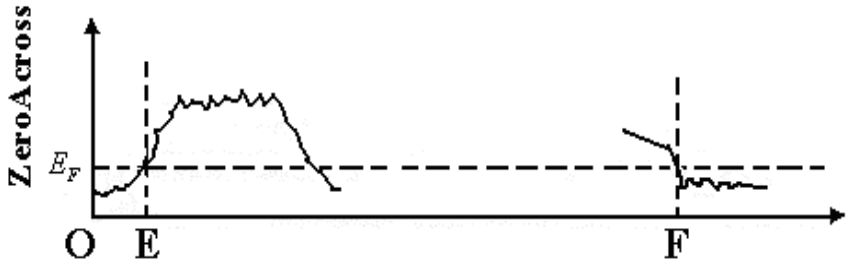

Figure 2. Second process of Dual-Threshold endpoint detection algorithm

Second process shown as Figure 2, a threshold (E_F) is set according to short-time average zero-crossing rate (Z_n) , scanning towards left from "C" can locate point "E", scanning towards right from "D" can locate point "F", "E-F" segment is the final detected result, the start-point of the speech signal is "E", the end-point is "F".

Dual-Threshold endpoint detection algorithm is simple and fast to speech signal which has high SNR, but it takes an assumption as prerequisite that there are some background noise frames at the beginning of voice signal, when background noise frames are not found or the background noise is powerful, the prerequisite may be breached, its detection result will be unreliable, which scope of application has some significant limitations.

In order to solve the above difficulties, this paper presents a new algorithm, which pre-processing is described briefly below.

a. Normalization on amplitude

The amplitude reflects intensity and relation of speech and background noise signal, in conventional practice scaling amplitude can't change characteristic of voice signal, therefore amplitude normalization can be executed by formula shown as follows.

$$y(i) = x(i) / \max(x(i)) \qquad (4)$$

In the equation, x(i) is the amplitude of original voice signal. Obviously, the amplitude of processed signal is unified, which scale will be in the interval [0,1], it will be beneficial to setting of threshold for subsequent processing.

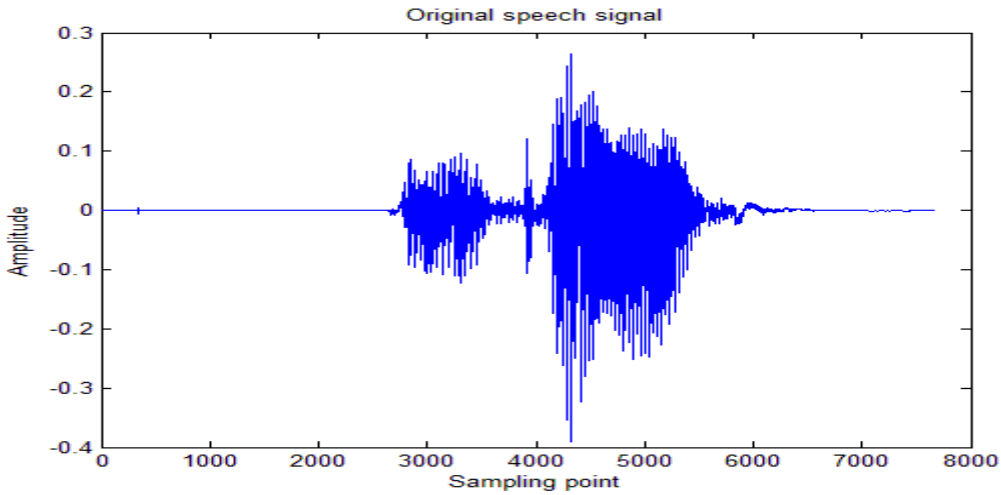

Figure 3. Original speech signal waveform

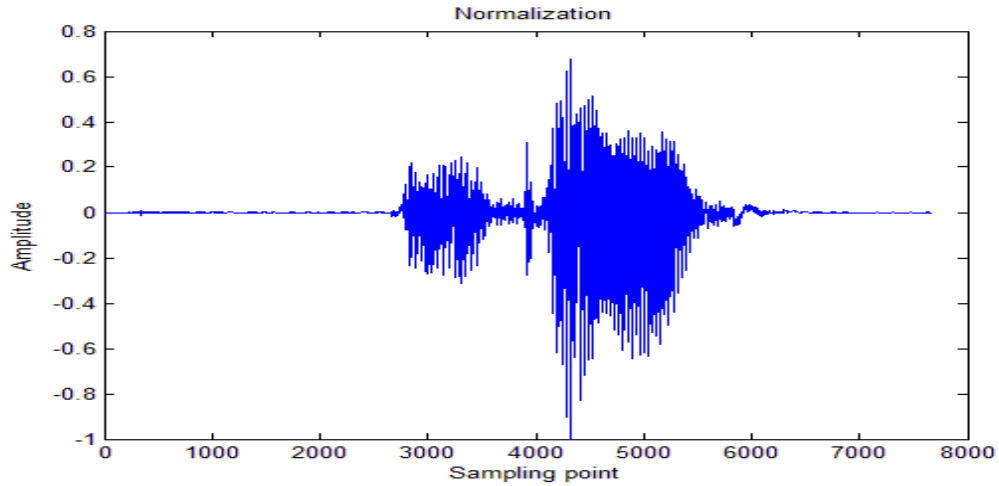

Figure 4. Normalization processing

b. Separating frame and calculating short-time amplitude

The sampling frequency of speech signal used in this paper is 8000 Hz, each data point represents 1/8000 second, that is 0.125ms. Because the characteristic of speech signal keeps mostly unchangeable in a short time range (10~30ms), the corresponding amount of data point is 80~240. Therefore, speech signal can be divided into some frames, each frame contains 80~240 data points, also called frame-length. There are some overlaps (frame-shift) between adjacent frames in order to keep smooth transition, frame-shift is half of frame-length in general.

To reduce computation load, short-time amplitude used as characteristics for detection in this paper, the formula is shown as the following.

$$STA(n) = \sum_{i=1}^{N} | y_n(i) | \qquad (5)$$

In the equation, y(i) is signal processed by normalization, N is frame-length, STA(n) is the short-time amplitude of frame n.

Figure 5 (frame-length is 256, frame-shift is 128) and Figure 6 (frame-length is 64, frame-shift is 32) are two waveforms of short-time amplitude, from which we can find that the waveforms of speech segment is greater than the waveforms of background noise segment obviously, this characteristic can be used for detection of speech endpoint. Comparing Figure 5 with Figure 6, it is inescapably clear that the bigger of frame-length and frame-shift, the less frames and its calculation amount, but its waveforms is more rough and the endpoint location is more fuzzy; the smaller of frame-length and frame-shift, the more frames and calculation amount, but its waveforms is more exquisite and the endpoint location is more accurate.

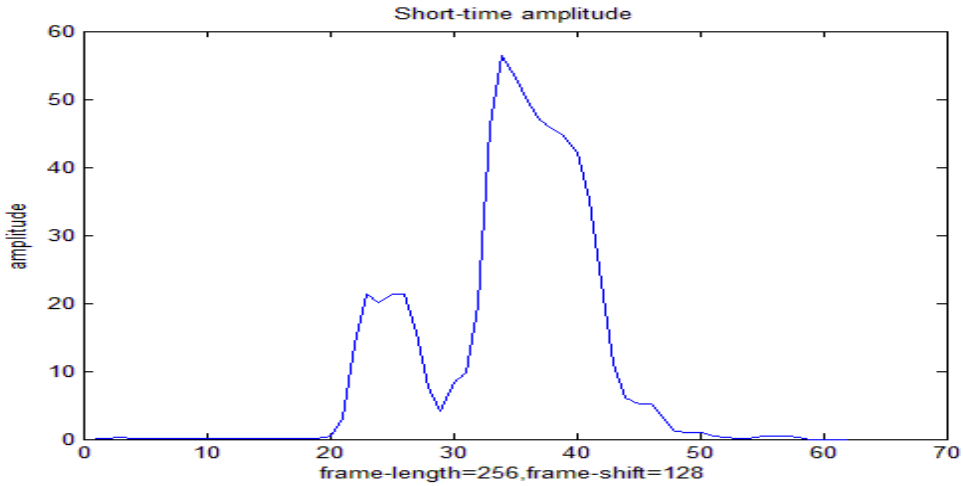

Figure 5. Short-time amplitude (256,128)

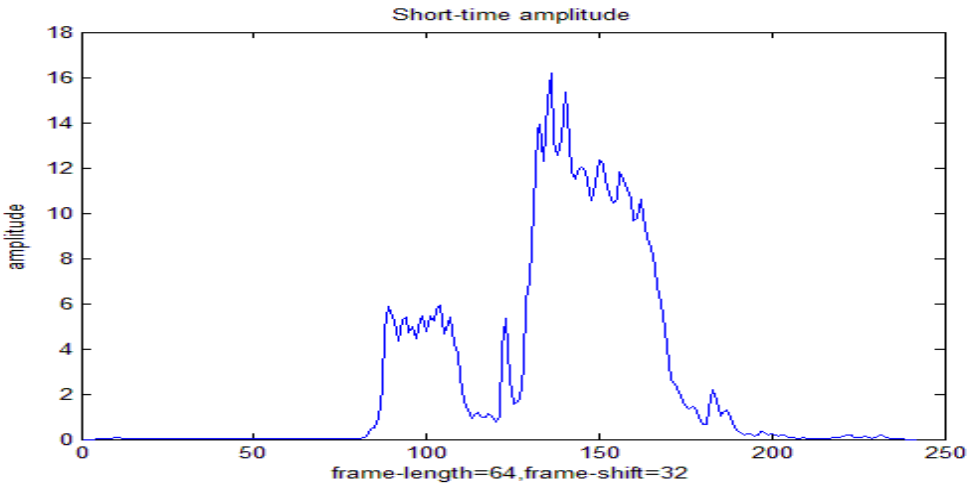

Figure 6. Short-time amplitude (64,32)

For ease of calculation, the frame-length of large-scale is 256, its time period is 32ms, the frame-length of small-scale is 64, their slot time are very close to the range of speech signal stationarity. Therefore, this paper use the large-scale(long frame-length and frame-shift) to scan whole speech signal, some endpoint (start-point and end-point) areas of effective speech segment can be rapid detected, this is rough detection; Then within these endpoint areas, small-scale(short frame-length and frame-shift) will be used for scanning forward or opposite direction according to the type of endpoint, which can finally obtain accurate start-point and end-point of effective speech segment, this is meticulous detection.

III. DESCRIPTION OF DETECTING PROCESS

a. Rough detection process

Rough detection aims to quickly find those areas contain start-point and end-point of effective speech segment, large-scale is used for scanning: frame-length is 256 and frame-shift is 128, which can reduce the amount of calculation, therefor can accelerate the speed of detection. Using large-scale also can avoid the interference of impulsive noise.

Through the analysis of the short-time amplitude waveform, it is easy to find that effective speech segment is peak shape, there is obvious difference between start-point area and end-point

area, therefore different methods were adopted to detect the start-point area and end-point area. The flowsheet of rough detection is shown in figure 7.

a.i Rough detection for start-point area

The waveform of start-point area is upward steep slope, the detection method is shown as below:

(1)Set the threshold of short-time amplitude to detect start-point area.

Because the data (amplitude of waveform) had been normalized into unified dimension, frame length is fixed (frame-length=256), therefore the threshold is easily determined, a large number of experiments show the optimum threshold is about 20 (Threshold_startarea=20);

(2)Read data by frame, calculate short-time amplitude: STA(i), i is frame sequence number, initialization i=1;

(3) If each STA(i) of three consecutive frames are greater than Threshold_startarea, then the start-point area is found (its location is the starting point of first frame), if not, continue to detect by frame.

Taking continuous three frames to judge threshold can avoid interference of the isolated noise effectively. Through the above processing, the endpoint areas of effective speech segment can be quickly found, and its location information can be recorded: Position_startarea=i.

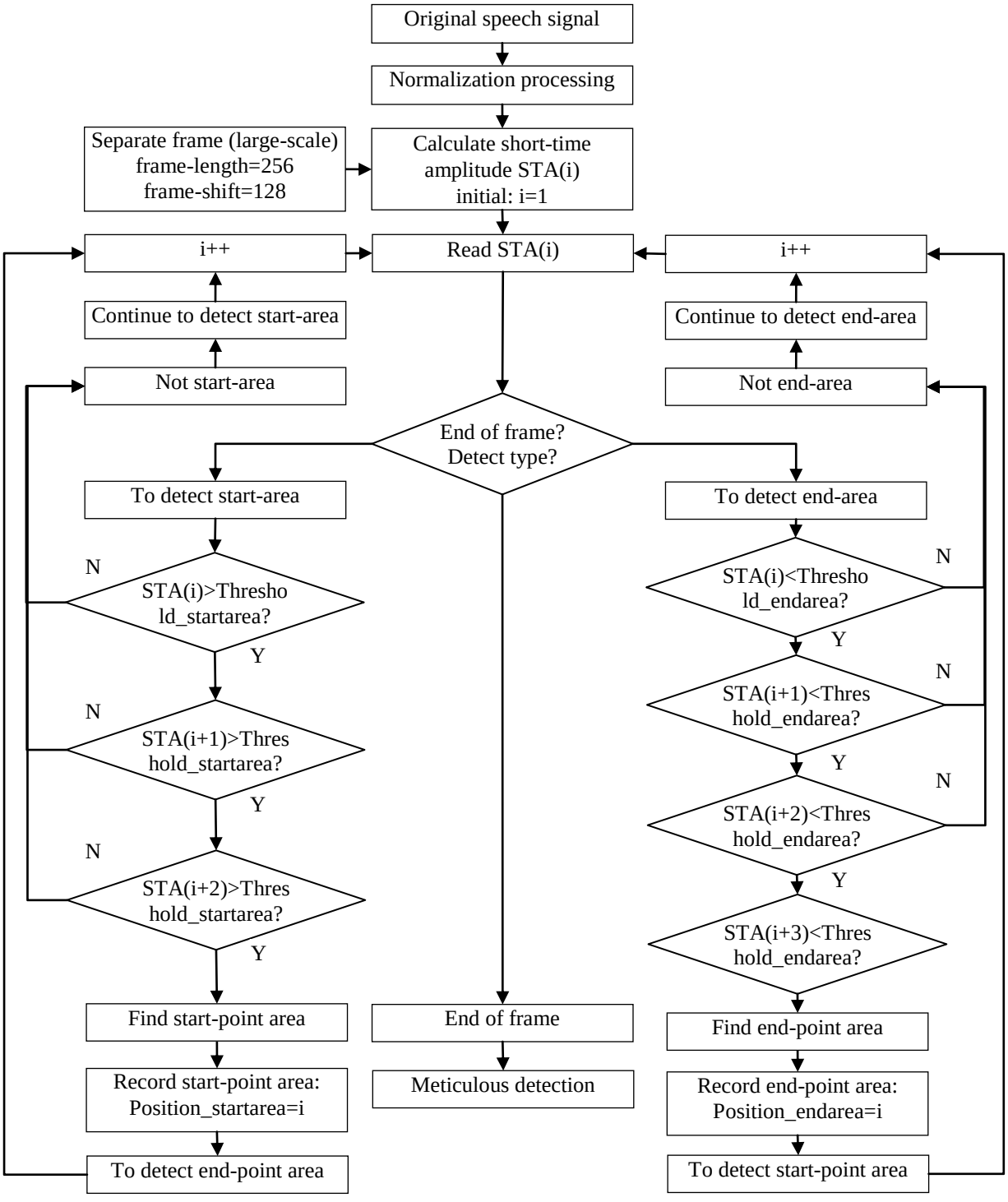

Figure 7. Flow chart of rough detection

a.ii Rough detection for end-point area

The waveform of end-point area is downward gentle slope, the detection method is shown as below:

(1) Set the threshold of short-time amplitude to detect end-point area.

Because the end sound of human voice is usually subdued, the threshold of end-point area should be set lower in order to ensure the integrity of end sound area. A large number of experiments show that the optimum threshold is about 12 (Threshold_endarea=12);

(2) Read data by frame, calculate short-time amplitude: STA(i), i is frame sequence number, initialization i=1;

(3) If each STA(i) of four consecutive frames are less than Threshold_endarea, then the end-point area is found (its location is the starting point of first frame), if not, continue to detect by frame.

Taking continuous four frames to judge threshold can avoid interference of the breath sound effectively. Through the above processing, the end-point area of effective speech segment can be quickly found, and its location information can be recorded: Position_endarea=i.

b. Meticulous detection process

Those areas obtained by rough detection are approximately region which contain start-point and end-point of speech segment, the aim of meticulous detection is to seek out the precise location of start-point and end-point. Because there is quite difference between start-point and end-point, different methods should be used for detection of start-point and end-point.

b.i Meticulous detection for start-point

The flowsheet is shown in figure 8.

(1) Selecting the scope to scan.

The start-point might well be located within small neighbor region radiated from start-point area which had been found by rough detection, three consecutive frames (frame-length=256) are selected as scanning scope for meticulous detection of start-point, which frame interval is [(Position_startarea-1)×256,(Position_startarea+1)×256], the little wide scope can effectively avoid missing detection or false detection of end-point;

(2) Setting scale and threshold of detection.

Using small-scale(short frame-length and frame-shift) to scan is advantageous to detect end-point accurately, when validated frame-length is 64, frame-shift is 32, the threshold is reduced

accordingly, a large number of experiments show the optimum threshold is about 5 (Threshold_startpoint=5);

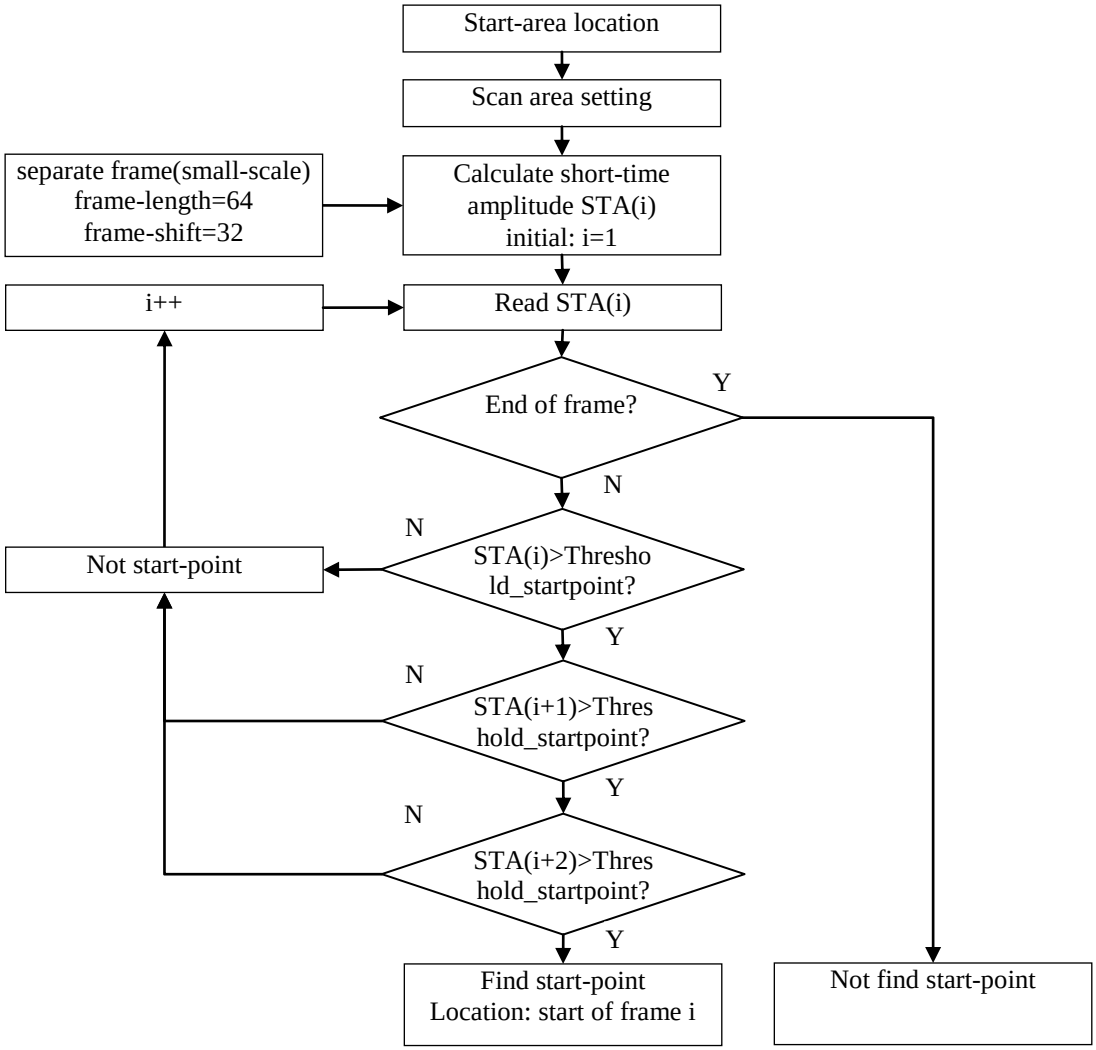

Figure 8. Flow chart of meticulous detection for start-point

(3) Read data by frame forward using small-scale, calculate short-time amplitude: STA(i);

(4) If each STA(i) of three consecutive frames are greater than Threshold_startpoint, then the start-point is found, its location is the starting point of first frame(frame i). If not, continue to detect by frame.

At last the start-point of effective speech segment can be found accurately.

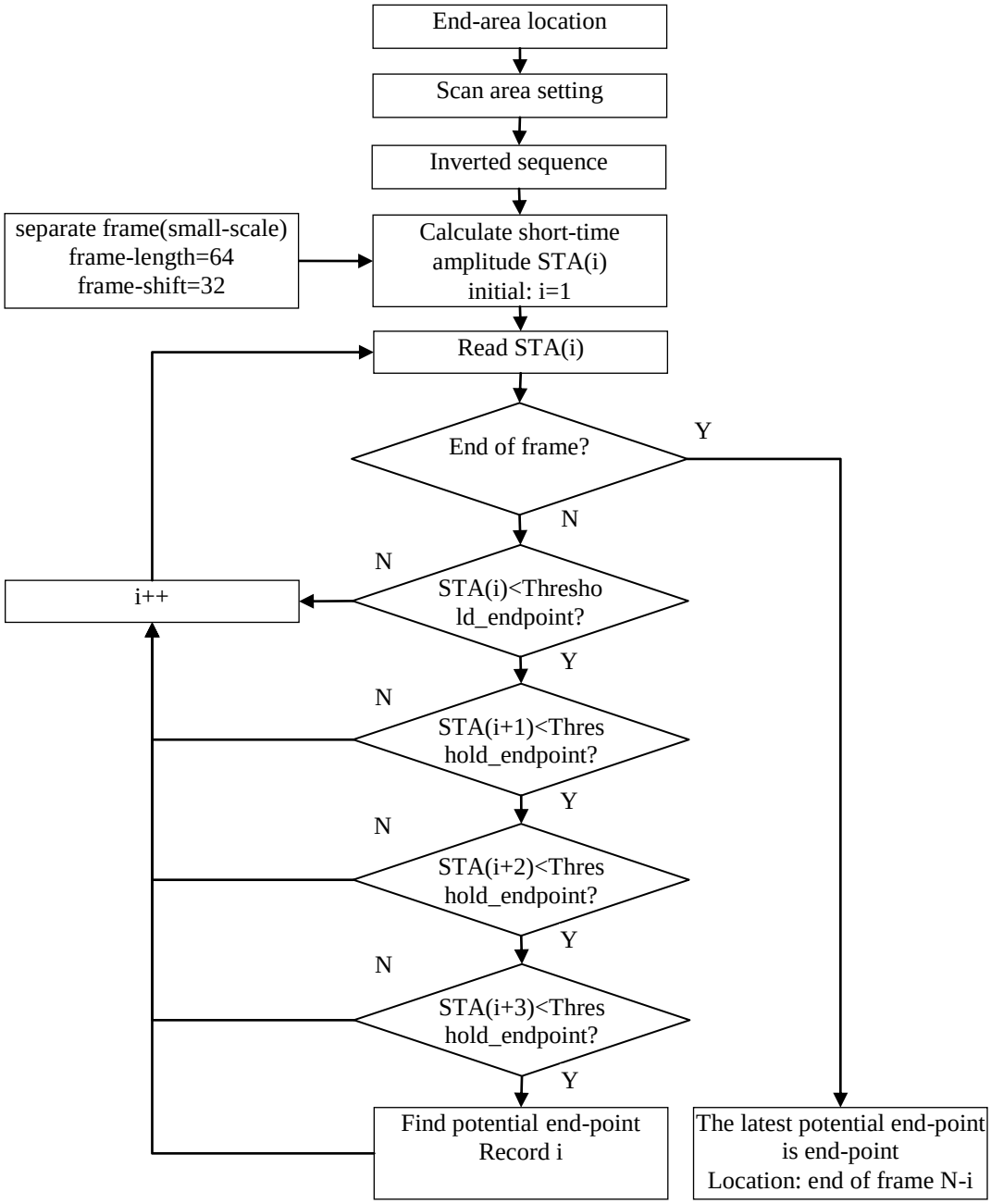

Figure 9. Flow chart of Meticulous detection for end-point.

b.ii Meticulous detection for end-point

The detecting method of end-point is quite different with start-point, the flowsheet is shown in figure 9.

(1) Selecting the scope to scan.

The end-point might well be located within small neighbor region originated from end-point area which had been found by rough detection, four consecutive frames (frame-length=256) are selected as scanning scope for meticulous detection of end-point, which frame interval is [Position_endarea×256,(Position_startarea+3)×256], the little wide scope can effectively avoid missing detection or false detection of end-point;

(2) Setting scale and threshold of detection.

As meticulous detection for start-point, same small-scale is used, the threshold validated by lot experiments is about 3 (Threshold_endpoint=3);

(3) Scanning in inverted sequence.

According to the feature of end-point waveform, backward scanning is better for detecting, thus the data of scanning region would be converted into inverted sequence;

(3) Read data by frame using small-scale, calculate short-time amplitude: STA(i);

(4) If each STA(i) of four consecutive frames are less than Threshold_endpoint, then the potential end-point is found, continue to scan until the end. If not, continue to detect by frame. At last, The latest potential end-point is true end-point,its location is the starting point of first frame(frame i in inverted sequence).

At last the end-point of effective speech segment can be found accurately.

Detecting process as shown in figure 10, the long vertical line(black dotted line) is the result of rough detection, the short vertical line(red real line) is the position of start-point and end-point after meticulous detection. It's easy to see from the waveform, accurate endpoint can be obtained by means of meticulous detection.

IV. EXPERIMENT RESULTS AND ANALYSIS

Voice sample tested in this paper consists of two part, one part is YOHO speech database collected by American ITT company, in which there are 1380 speech samples which content is English digits, 100 samples were selected to test and another 50 samples were selected to train from YOHO database; Another part is a small speech database recorded by ourselves, which include 20 persons (11 male voice and 9 female voice), each one had spoken 10 voice segments which content is Chinese phrase, from which 50 samples were selected to test and another 20

samples were selected to train. Those samples were recorded indoor environment, sampling rate is 8000Hz, 16 bits data, 128Kbps.

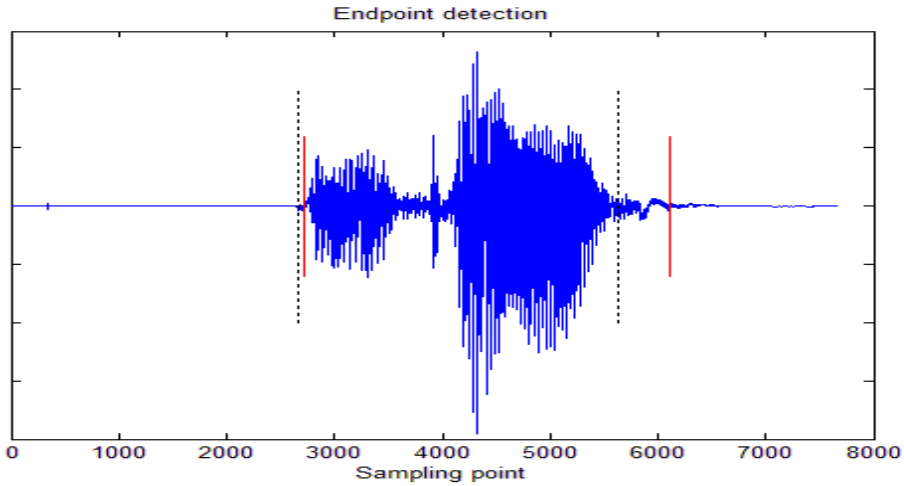

Figure 10. Rough and meticulous detection

a. Effectiveness analysis of endpoint detection

All the samples that were selected were tested for endpoint detect, experiments show positive results, nearly all of endpoint can be detected precisely and fleetly.

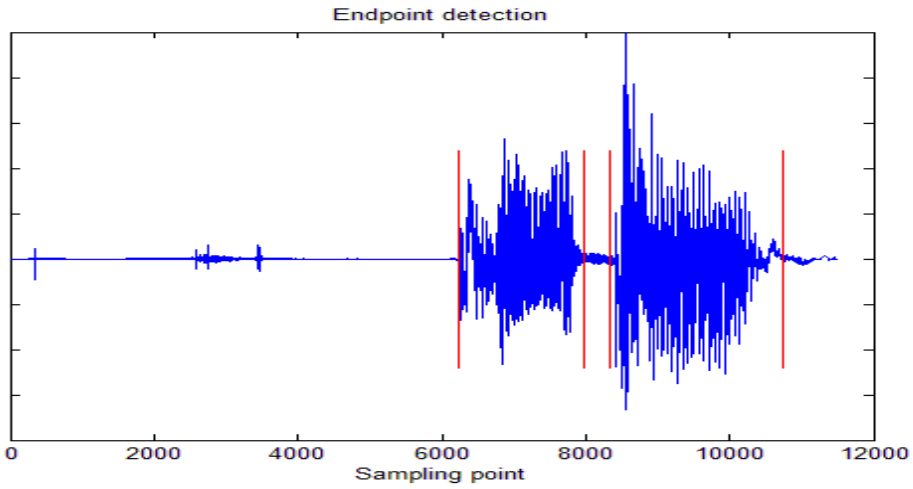

Figure 11. The detection results to noisy signal

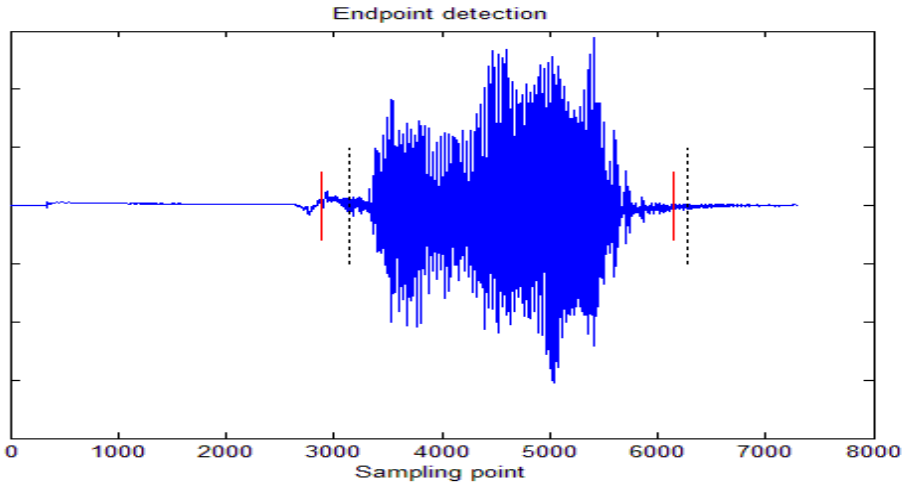

Figure 12. Comparison of test results

In figure 11, there is some isolated noise in speech signal, the algorithm of this paper can filter out those impulsive noise or background noise, find correct position of speech endpoint.

The result of comparing this algorithm with Dual-Threshold detection method is shown in figure 12, the long vertical line (black dotted line) is the position detected by Dual-Threshold detection method, the short vertical line (red real line) is the position detected by the algorithm of this paper. Obviously, endpoint location detected by this paper is more accurate. In addition, Dual-Threshold detection method deems there should be silence segment in the beginning of voice signal, the assumption is not always true, missing detection or false detection maybe occur in some cases. Experiments show that the algorithm of this paper is robust.

b. Effectiveness analysis of speech recognition

In order to inspect the influence on speech recognition from endpoint detection, experiment of speaker recognition was tested based on two method of endpoint detection, which adopt GMM (Gaussian mixture model) algorithm, feature parameter is MFCC coefficient(16 dimension), GMM order is 32. The test result is compared in Table 1.

The experimental shows that detecting speed of this method is faster, an extra reason is that this method omit the calculation on short-time zero crossing rate, which can accelerate detection speed. Because this algorithm can get more accurate endpoint, therefore it can improve recognition performance effectively.

Table 1: The statistical data of experiment

Samples		**Dual-Threshold method**	**method of this paper**	**compare**
YOHO samples	amount	100	100	-
	length(ms)	3962	3962	-
	detect time (ms)	172	127	- 26.2%
	recognition rate(GMM)	92.6%	97.3%	+ 4.7%
Chinese words samples	amount	50	50	-
	length(ms)	1126	1126	-
	detect time (ms)	119	86	-27.7%
	recognition rate(GMM)	90.6%	96.5%	+ 5.9%

V. CONCLUSIONS

According to the characteristics of speech signal, this paper put forward a new method for endpoint detection based on a major improvement to Dual-Threshold detection algorithm, which set short-time amplitude as characteristic, large-scale (long frame-length and frame-shift) was used for rough detection, then start-point and end-point areas can be found quickly, from which smaller scanning regions can be screened out for meticulous detection, small-scale (short frame-length and frame-shift) is used to scan those regions forward or reversely, the accurate position of start-point and end-point of speech signal can be obtained finally.

This paper adopted two kind scale, large-scale determines the speed of detection, small-scale decides the accuracy of detection, both can be adjusted to satisfy the different requirements (detection threshold can also be adaptive adjustment accordingly), the method of this paper ensures both detection speed and precision, which has more flexibility and applicability.

Experiment results show that the method of this paper can detect endpoints of voice signal more quickly and accurately, which robustness is good, it can satisfy the needs of endpoint detection

under the common environment. Because the endpoint detected by the method of this paper is more precise, it can improve recognition performance effectively.

REFERENCES

[1] Savoji M H. A robust algorithm for accurate endpointing of speech signals[J]. Speech Communication, 1989, 8(1): 45-60.

[2] L.R.Rabiner, B.H. Juang. Fundaments of Speech Recognition[M], PrentieeHall,1993.

[3] Shen Yaqiang. Voice activity detection algorithm with low signal-to-noise based short-time fractal dimension of signals[J].Chinese Journal of Scientific Instrument, 2006.6(27):2310~2312.

[4] HU Guang-rui,WEI Xiao-dong. Endpoint detection of noisy speech based on cepstrum[J]. Acta Electronica Sinica, 2000, 28(10):95~97.

[5] Shen Jialin, Huang Jeihweih, Lee Linshan. Robust entropy-based endpoint detection for speech recognition in noisy environments[C] //Proc of ICSLP 98. Sydney: Australian Speech Science and Technology Association Incorporated, 1998:232~235.

[6] Huang Liangsheng, Yang Chungho. A novel approach to robust speech endpoint detection in car environments[C] //Proc of ICASSP 00. Piscataway, NJ: IEEE, 2000: 1751-1754.

[7] LI Ru-wei,BAOA Chang-chun. Speech EndPoint Detection Algorithm Based on the Band-Partitioning Spectral Entropy and Spectral Energy[J], Journal of Beijing University of Technology, 2007(9):920-924.

[8] Zhao Huan, Zhao Lixia, Zhao Kai, et al. Voice activity detection based on distance entropy in noisy environment [C] //Proc of the 5th Int Joint Conf on INC, IMS and IDC. Los Alamitos, CA: IEEE Computer Society, 2009: 1364-1367.

[9] TIAN Ye. Robust word boundary detection through linear mapping of the sub-band energy in noisy environments[J], Journal of Tsinghua University (Sci &Tech), 2002; 42(7); 953-956.

[10] LIU Hong-xing, DAIBei-qian, LU Wei.A Speech Endpoint Detection Method Based on Consonance Energy[J], Computer Simulation,2008,11(25):305-308.

[11] C Bandt,B Pompe. Permutation entropy: a natural complexity measure for time series [J]. Physical Review Letters, 2002, 88(17): 174102-1-4.

[12] Fan Yingle, Wu Chuanyan, Li Yi, et al. Application of C0 complexity measure in detecting speech [J]. Chinese Journal of Sensors and Actuators, 2006, 19 (3): 750-753.

[13] SHI Wei,ZOU Yue-xian. Voice Activity Detection Algorithm with Low Signal-to-Noise Ratio Based on Hilbert-Huang Transform[J],Technical Acoustics,2011,12(30):281-282.

[14] Wang Ming-he,Zhang Er-hua,Tang Zhen-min,et al. Voice Activity Detection Based on Fisher Linear Discriminant Analysis[J]. Journal of Electronics & Information Technology, 2015,37(6):1343-1349.

[15] Xiao-Lei Zhang, Ji Wu. Deep belief networks based voice activity detection[C]. IEEE Transactions on Audio, Speech, and Language Processing, 2013,21(4):697-710.

[16] ZHU heng-Jun,YU Hong-bo,WANC1 Fa-zhi. Speech Endpoints Detection Algorithm Based on Support Vector Machine and Wavelet Analysis[J]. Computer Science,2012,39(6):244-265.

[17] Ryant N, Liberman M, Yuan Jia-hong. Speech activity detection on YouTube using deep neural networks[C]. Interspeech: 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013: 728-731.

[18]Kim Dong Kook, Shin Jong Won, Chang Joon-Hyuk. Enhanced voice activity detection in kernel subspace domain[J]. The Journal of the Acoustical Society of America, 2013,134 (1):EL70-6.

[19] A.M. Aibinu, M.J.E.Salami, A.A. Shafie. Artificial neural network based autoregressive modeling technique with application in voice activity detection[J]. Engineering Applications of Artificial Intelligence, 2012, 25 (6):1265-1276.

[20]Kim Dong Kook, Chang Joon-Hyuk. Statistical voice activity detection in kernel space[J]. Journal of Acoustical Society of America, 2012, 132 (4):EL303-9.

[21] Kun-Ching Wang. Voice Activity Detector for Noise Spectrum Estimation Using a Dynamic Band-Splitting Entropy Estimate [J]. International Journal of Computers and Applications, 2011, 33 (3):220-228.

[22] Jinsoo Park, Wooil Kim, David K.Han,et al. Voice Activity Detection in Noisy Environments Based on Double-Combined Fourier Transform and Line Fitting[J]. The Scientific World Journal, 2014, Vol.2014.

[23] Sang-Yeob Oh, Kyungyong Chung. Improvement of Speech Detection Using ERB Feature Extraction[J]. Wireless Personal Communications, 2014, 79 (4):2439-2451.

[24]CHAO Hao,YANG Zhan-lei,LIU Wen-ju. Itegrating articulatory information into stochastic segment models for continuous Mandarin speech recognition[J].Application Research of Computers,2014,31(11):3365-3368.

[25] Shweta Sinha, Aruna Jain ,S. S. Agrawal. Acoustic-phonetic feature based dialect identification in Hindi speech[J]. International Journal On Smart Sensing and Intelligent Systems.2015,8(1):237-254.