TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 13, No. 1, January 2015, pp. 137 ~ 144
DOI: 10.11591/telkomnika.v13i1.6902

Voice Activity Robust Detection of Noisy Speech in Toeplitz

Jingfang Wang
School of Information Science and Engineering, Hunan International Economics University,
Changsha, China, postcode:410205
email: matlab_bysj@126.com

Abstract

A Toeplitz de-noising method using the maximum eigenvalue is proposed for the voice activity detection at low SNR scenarios. This method uses the self-correlation sequence of speech bandwidth spectrum to construct a new symmetric Toeplitz matrix and to compute the largest eigenvalue, and the double decision thresholds in the largest eigenvalue are applied in the decision framewok. Simulation results show that the presented algorithm is more effective in distinguishing speech from noise and has better robustness under various noisy environments. Compared with novel method of recurrence rate analysis, this algorithm shows lower wrong decision rate. The algorithm is of low computational complexity and is simple in real-time realization.

***Keywords**: voice activity detection (VAD), speech bandwidth spectrum, maximum eigenvaluem, robustness, Toeplitz matrix*

1. Introduction

Voice is an acoustic performance of language, it is auditory organ perception on the mechanical vibration of outside sound propagation medium, and it is an important carrier of human information transmission and emotional communication. Currently, voice processing technologies require voice input in a quiet environment, when there is the ambient noise (such as factories, airports, etc.), system performance decreases dramatically. However, voice communication process would inevitably be affected from the surrounding environment noise, such as propagation medium. Voice activity detection is an important part of the digital speech processing [1-5], the aim is to detect the speech signal segments and the noise segments from the sampled digital signal. The collection of voice signal is divided into pure noise and noisy speech segment, the beginning and ending points of the speech segment is determined, speech endpoint test is an important part of the speech enhancement algorithms and speech coding. In the speech recognition process, if the beginning and ending of the speech segment endpoints are correctly determined, and the amount of computation and the error rate of speech recognition can be reduced.

In endpoint detection algorithm, short-term energy is the most common features [6], which can effectively separate the speech and noise at high SNR environment, but a large number of experimental results show that short-term energy approach performance has declined markedly at low SNR environment and non-stationary noise environments. Of course, part of the algorithm can maintain stable performance in the low SNR environment [7]. The disadvantage is that the computational complexity is too large; it is not suitable for real-time speech recognition system. Shen [8] first proposed the entropy for speech / noise classification, and the noise difference in the person's speech can be expressed from their spectral entropy. Speech spectrum entropy algorithm is better at low SNR environment than approach based on energy. It is better in the white noise, but it still difficult to work in colored noise.

The signal subspace [9-12] has be used in terms of speech enhancement. Because it is difficult to achieve voice endpoint detection at low SNR and non-stationary noise conditions, a de-noising speech endpoint detection method is proposed based on Toeplitz largest eigenvalues in this paper. In this method, speech spectrum autocorrelation sequence is used to construct a symmetric Toeplitz matrix, the information amount of the matrix maximum eigenvalues is used, and the speech signal endpoints are detected by the double threshold. The

Received September 11, 2014; Revised October 23, 2014; Accepted November 15, 2014

algorithm greatly improves the VAD detection accuracy and validity, the algorithm can be able to maintain a better detection performance in a variety of noisy environments and low SNR conditions.

2. Structure of Toeplitz Information Matrix

Voice signal and characterization of their properties change randomly on the whole, and its parameters are the essential characteristics change over time, and it is a typical non-stationary process, but in a short period of time (10 ~ 30ms), its properties remain relatively stable, and therefore it can be seen as a quasi-stationary process, that is short-term stability of the speech signal. At present, most of the speech signal processing techniques are in based on the "short-time", the voice signal is divided into a plurality of segments, and its characterized parameters are analyzed, where each section is called a "frame", the process of segments is said as "framing" process, it is to be achieved through the voice signal windowing function, frame size generally takes 10 ~ 30ms. Sub-frames can be a continuous segment, but it is generally carried out by overlapping segments of a sliding window, such that the smooth transition can be done between frames, which can maintain the continuity of the signal. On selected window function, in order to get a high frequency resolution and overcome Gibbs phenomenon, we choose Hanning (Hanning) window in overlap style segments.

Noisy speech signal x (n) is framed, frame size is FrameLen, frame shift is StepLen (StepLen <FrameLen), the total number of frames is Num, if fast Fourier transform (FFT) is made to the k-th frame signal, we get the spectrum NFFT points YF (i, k) (0 ≤ i ≤NFFT), speech frequency range is between 200Hz and 4kHz, the corresponding point of interval [Nd, Ng] point (0≤Nd<Ng≤NFFT) is obtained, L= Ng-Nd+1, LM=L/2 is the size of the Toeplitz matrix; Xk(i)= YF(i+Ng-1,k) (1≤i≤L).

K-th frame autocorrelation sequence of speech spectrum is R (m):

$$R(m) = \frac{1}{L-m}\sum_{i=1}^{L-m} X_k(i)X_k(i+m), m = 0,1,2,...,LM-1 \tag{1}$$

LM-dimensional structure of real symmetric Toeplitz matrix A:

$$A = Toeplitz(R) = \begin{bmatrix} R(0) & R(1) & \bullet\bullet\bullet & R(LM-1) \\ R(1) & R(0) & \bullet\bullet\bullet & R(LM-2) \\ \bullet\bullet\bullet & \bullet\bullet\bullet & \bullet\bullet\bullet & \bullet\bullet\bullet \\ R(LM-1) & & \bullet\bullet\bullet & R(0) \end{bmatrix} \tag{2}$$

This order of Toeplitz matrix is not high, seeking eigenvalue at fast speed.

3. Realization of Voice Activity Detection

3.1. Iterative Method for Maximum Principle of Eigenvalue

Matrix power method is to strive for the largest eigenvalue and corresponding eigenvector of an iterative method. An n-set are linearly related to the eigenvectors v1, v2, …, vn, the corresponding eigenvalue λ1, λ2, …, λn satisfy:

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| \tag{3}$$

3.1.1. The Basic Idea

Because $\{v_1, v_2, \ldots, v_n\}$ is a basis of C^n, So any given $x^{(0)} \neq 0$, $x^{(0)} = \sum_{i=1}^{n} a_i v_i$'s linear representation.

$$\begin{aligned} A^k x^{(0)} &= A^k(\sum_{i=1}^{n} a_i v_i) = \sum_{i=1}^{n} a_i A^k v_i \\ &= \sum_{i=1}^{n} a_i \lambda_i^k v_k = \lambda_1^k [a_1 v_1 + \sum_{i=2}^{n} (\frac{\lambda_i}{\lambda_1})^k a_i v_i] \end{aligned} \tag{4}$$

If $a_1 \neq 0$,then $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$, when k is arge enough, $A^{(k)}x^{(0)} \approx \lambda_1{}^k a_1 v_1 = cv_1$ is λ_1's eigenvectors.

On the other hand, max(x) = x_i, where $|x_i| = ||x||_\infty$, When k is sufficiently large,

$$\frac{\max(A^k x^{(0)})}{\max(A^{k-1}x^{(0)})} \approx \frac{\max(\lambda_1{}^k a_1 v_1)}{\max(\lambda_1{}^{k-1} a_1 v_1)} = \frac{\lambda_1{}^k \max(a_1 v_1)}{\lambda_1{}^{k-1} \max(a_1 v_1)} = \lambda_1$$

If a1 = 0, due to rounding error, and there will be an iteration vector component in the v1 direction is not 0, iteration continues λ1 can be obtained and the corresponding approximate eigenvectors.

3.1.2. Standardization

In practical calculation,if λ1| > 1,|λ1ka1| →∞;or if |λ1| < 1,| λ1ka1| → 0will stop. A "standardized" approach.

$$\begin{cases} y^{(k)} = \dfrac{x^{(k)}}{\max(x^{(k)})} \\ x^{(k+1)} = Ay^{(k)} \end{cases} \quad k = 0,1,2,\ldots \tag{5}$$

Theorem: Given any initial vector $x^{(0)} \neq 0$,

$$\begin{cases} \lim\limits_{k\to\infty} y^{(k)} = \dfrac{v_1}{\max(v_1)} \; Eigenvecto \;\; rs \\ \lim\limits_{k\to\infty} \max(x^{(k)}) = \lambda_1 \; Eigenvalue \end{cases} \tag{6}$$

Proof:

$$\begin{aligned} y^{(k)} &= \frac{x^{(k)}}{\max(x^{(k)})} = \frac{Ay^{(k-1)}}{\max(Ay^{(k-1)})} \\ &= \frac{A\dfrac{x^{(k-1)}}{\max(x^{(k-1)})}}{\max(A\dfrac{x^{(k-1)}}{\max(x^{(k-1)})})} = \cdots = \frac{A^k x^{(0)}}{\max(A^k x^{(0)})} \\ &\overset{(4)}{=} \frac{\lambda_1{}^k [a_1 v_1 + \sum_{i=2}^{n} a_i (\frac{\lambda_i}{\lambda_1})^k v_i]}{\max\left\{\lambda_1{}^k [a_1 v_1 + \sum_{i=2}^{n} a_i (\frac{\lambda_i}{\lambda_1})^k v_i]\right\}} \end{aligned}$$

$$= \frac{[a_1 v_1 + \sum_{i=2}^{n} a_i (\frac{\lambda_i}{\lambda_1})^k v_i]}{\max\left\{[a_1 v_1 + \sum_{i=2}^{n} a_i (\frac{\lambda_i}{\lambda_1})^k v_i]\right\}} \xrightarrow{k\to\infty} \frac{a_1 v_1}{\max(a_1 v_1)} = \frac{v_1}{\max(v_1)}$$

$$\max(\ x^{(k)}) = \max(\ Ay^{(k-1)})$$

$$\xrightarrow{k\to\infty} \max(\ A\frac{v_1}{\max(\ v_1)}) = \frac{\max(\ Av_1)}{\max(\ v_1)}$$

$$= \frac{\max(\ \lambda v_1)}{\max(\ v_1)} = \lambda_1$$

Note: If the eigenvalues does not satisfy condition (3), the power law convergence of the analysis is λ1 = λ2 = … = λ r complicated, but if λ1| > |λ r +1| ≥ …≥ |λn| is Theorem conclusion stillλ holds. At this time iteration with different initial vector vector sequence generally tend to feature vectors of λ1 different.

3.2. The Largest Eigenvalue Algorithm for Toeplitz Matrix A

In order to solve one of the largest eigenvalue, where we use the power method, which can avoid the matrix decomposition or inverse matrix calculations in seeking eigenvalues. The implementation steps:

1) Initial values:LM-dimensional column vector y=[1,1,…,1]^H, H is transpose; LM-dimensional column vector y_0=[0,0,…,0]^H; Decision cycle conditions eps=0.0001(A smaller number), d=1.

2) Matrix:z=Ay

3) Normalized:

$$y = \frac{z}{\|z\|_\infty}, \quad \|z\|_\infty = \max\{|z(i)|, i = 1,2,...,LM\} \tag{7}$$

4) Calculation: $d = \max\{|\ y(i) - y0(i)\ |, i = 1,2,..., LM\}$ The last y is reserved, y0=y

5) Cycle verdict: If d> eps, to turn (2) step, or to turn (6) step.

6) To calculate the largest eigenvalue:

$$\lambda = \max\{|\ z(i)\ |, i = 1,2,..., LM\} \tag{8}$$

7) To retain the largest eigenvalue information in k-th frame:

$$Tzv(k) = 10\log_{10}(\lambda) \tag{9}$$

3.3. Double threshold Voice Endpoint Discrimination

In order to prevent that the largest eigenvalue information Tzv appear jagged fluctuations between frames, average filtering is done in the Tzv of adjacent three frames. Double threshold Voice Endpoint Discrimination step:

Step 1: to identify the initial frame N0 as the noise frame, to calculate the average value Avg the standard deviation Std of Tzv(l)(0<l<= N_0) . Double threshold is defined as threshold TS in speech frames and TN thresholds in noise frames, they are in Formula (10):

$$\begin{aligned} &TN=Avg+\alpha{*}Std \quad \alpha>0 \\ &TS=Avg+\beta{*}Std, \quad \beta>\alpha \end{aligned} \tag{10}$$

Step 2: to calculate the largest eigenvalue information Tzv (l) of the next frame speech signal. When the up frame is the noise frame, if the Tzv(l)<TS, l-th frame is as noise frames, Tzv(l) of speech frames was greater than TS. When the up frame is the speech frame, Tzv(l) is

compared with threshold TN, if the Tzv(l)<TN, l-th frame is as the noise frames, Tzv(l) of speech frames was greater than TN. Signal sampling loops up to the end of step 2.

α, β can be selected between the (0,4), different value α, β are selected at different noise. The speech segment continues at least some time, such as 0.2 seconds. If the detected speech segment is less than Tzv (l), it is called "Voice debris" (in the non-Gaussian noise [eg: factory noise (facyory), loud noise (babble)] under common), the last of isolation "Voice debris" is removed or on the adjacent "Voice debris" is integration.

4. Experimental Evaluation

Background noise is taken from Noisex-92 database [13], and its sampling frequency fs = 19.98kHZ. We have the same sampling frequency fs, the noise in the computer record and interior noise environment, "language, tone, end point" sound shown in Figure 1(a), the method frame line for the endpoint detection results. Process in the voice sub-frames, each frame taking 25ms, the frame size FrameLen = [0.025fs] point, frame shift$[\frac{FrameLan}{4}]$, to determine the fast Fourier transform of each frame (FFT) length of the take is equal to frame length FrameLen, interception tarted the noise frame N_0=20.

The original voice, original voice and noise Noisex-92 library noise - white noise (white), pink noise (pink), aircraft noise (f16_cockpit), were loud noise (babble) noise Toeplitz matrix with the largest eigenvalue article endpoint detection method, the signal to noise ratio SNR = 5dB, 0dB,-5dB, the use of recursive algorithm and signal analysis method [14] compared the test results are presented in Figure 1-3. Left part of the figure the abscissa is time (seconds), the vertical axis for the range; the middle of the abscissa is the number of frames, the vertical axis is the largest eigenvalue Toeplitz matrix information (dB); right side of the abscissa is the number of frames, longitudinal coordinates for the recursive degrees (%). Figure SNR = 5dB in the left part of S, for voice, speech mixed with different noise and their detection, the central figure of the Toeplitz matrix algorithm for the largest eigenvalue divided line and endpoint information; this algorithm in various noise mixing Next, Toeplitz matrix maximum eigenvalue curve is not the amount of information, voice endpoint segmentation accuracy, good adaptability.

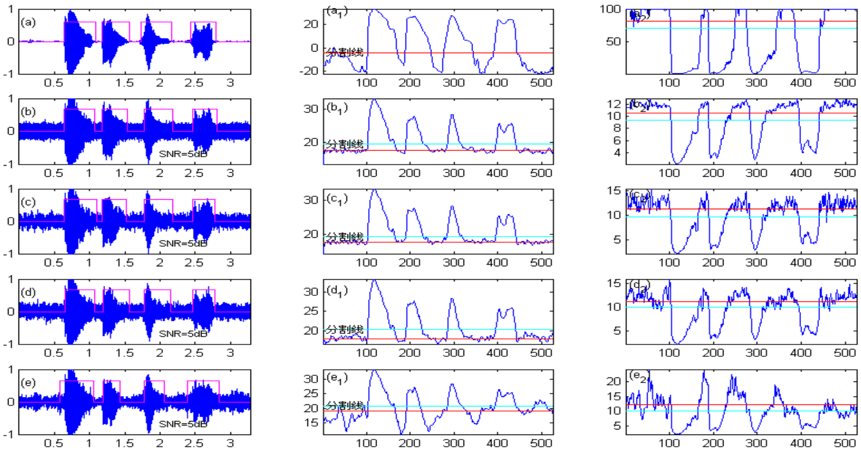

Figure 1. The original voice and mixed with different noise (SNR = 5dB) compared Endpoint Detection

(a) The original speech endpoint detection and the endpoint detection algorithm, (a1) of the original voice algorithm information Toeplitz matrix eigenvalue curve and the maximum partition line; (a2) corresponding to the measure of the signal curve of recursive analysis and segmentation;

(b) Mixed White Noise (white) and the speech endpoint detection algorithm, (b1) Mixed plant noise, the algorithm information Toeplitz matrix eigenvalue curve and the maximum partition line; (b22) the corresponding measure of the signal curve of recursive analysis and segmentation;

(c) Mixture of pink noise (pink) Voice of the endpoint detection algorithm, (c1) mixed pink noise algorithm information Toeplitz matrix eigenvalue curve and the maximum partition line; (c2) corresponds to recursive analysis of the measurement signal curve segmentation;

(d) Mixed fighter cockpit (f16_cockpit) and the noise of the speech endpoint detection algorithm, (d1) hybrid aircraft noise algorithm information Toeplitz matrix eigenvalue curve and the maximum partition line; (d2) corresponding to the measure of the signal curve of recursive analysis and segmentation;

(e) Loud noise mixed people (babble) and the algorithm of speech endpoint detection, (e1) loud noise hybrid algorithm were the largest eigenvalue Toeplitz matrix with the dividing line information curve; (e2) corresponds to recursive analysis of the measurement signal curve and segmentation.

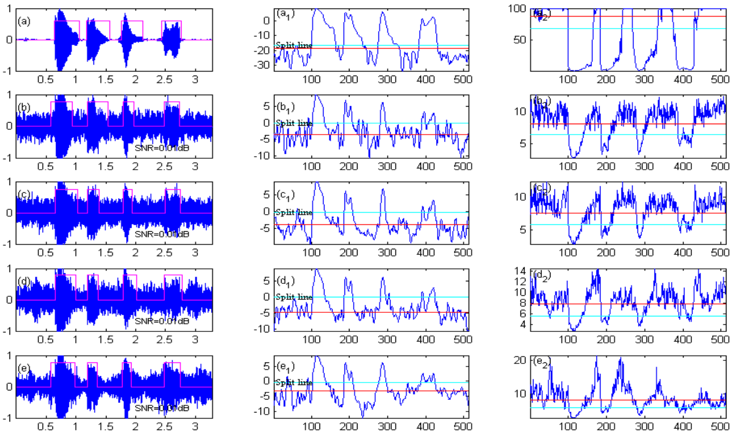

Figure 2. The original voice and mixed with different noise (SNR = 0dB) of the endpoint detection contrast (similar to Figure 1 legend)

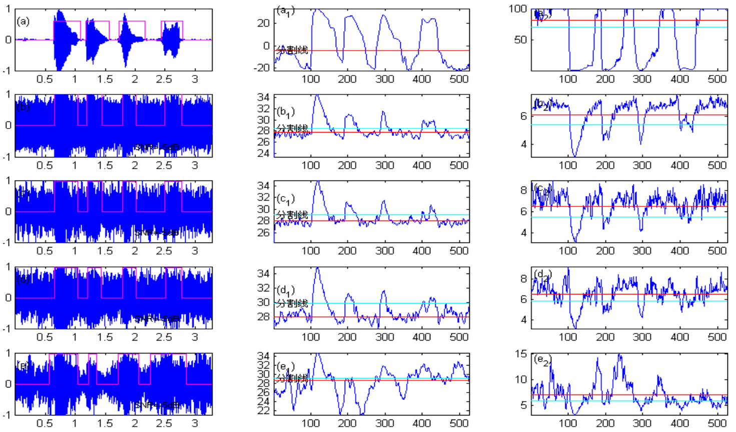

Figure 3. The original voice and mixed with different noise (SNR =- 5dB) compared Endpoint Detection

To further evaluate the algorithm performance, quantitative analysis of the merits of the algorithm, this paper select the following three indicators to measure [15-17]:

$$P_{SA}=1-\frac{N_{1,0}}{N_1},P_{NA}=1-\frac{N_{0,1}}{N_0},$$

$$P_A=1-\frac{N_{0,1}+N_{1,0}}{N_0+N_1} \tag{11}$$

Where, N_1 and N_0 were hand-labeled test speech in voice frames and the total number of noise frames,$N_{1,0}$ for the hand-labeled speech frames and error frames for noise identification number, $N_{0,1}$ frame for the hand-marked and identified as noise wrong number of speech frames. Then P (A / S) is the correct rate of detection of speech frames, P (A / N) to detect the correct frame rate of the noise, P (A) of the total detection accuracy. Table 1 shows the different noise in different SNR environment the two methods results summary table.

Table 1. Endpoint detection test results

	Detector	ToeplitzMaximum eigenvalue algorithm			Signal recursive algorithm		
Noise	Correct rate (%)	P(A/S)	P(A/N)	P(A)	P(A/S)	P(A/N)	P(A)
white	SNR=5dB	91.90	97.48	94.86	79.35	100.00	90.29
	SNR=0dB	80.57	100.00	90.86	68.02	100.00	84.95
	SNR=-5dB	68.83	100.00	85.33	51.01	100.00	76.95
pink	SNR=5dB	91.90	97.48	94.86	78.95	100.00	90.10
	SNR=0dB	79.76	100.00	90.48	67.61	100.00	84.76
	SNR=-5dB	69.23	100.00	85.52	47.37	100.00	75.24
f16	SNR=5dB	91.90	100.00	96.19	81.38	78.06	79.62
	SNR=0dB	72.47	100.00	87.05	71.66	77.70	74.86
	SNR=-5dB	66.80	100.00	84.38	63.56	77.70	71.05
babble	SNR=5dB	78.54	77.34	77.90	87.45	50.72	68.00
	SNR=0dB	73.58	77.34	75.62	80.97	50.72	64.95
	SNR=-5dB	74.90	60.43	67.24	72.47	50.72	60.95

5. Conclusions and Outlook

A new method of robust noisy speech endpoint detection is proposed from new visual angle in this paper, which is based on the maximum eigenvalue of Toeplitz; In this method, autocorrelation sequence with the spectral range (200Hz - 4kHz) is used to construct a symmetric Toeplitz matrix, the maximum eigenvalues of the matrix is used on the endpoint detection of the speech signal pairs threshold. Main signal is extracted by using the maximum eigenvalue, and noise is suppressed better. When the SNR is below 5dB, the general endpoint detection methods seem almost powerless, such as short-time spectral estimation, this algorithm is still useful, it has to calculate simple, noise immunity characteristics is strong, and experiments show that the method of correctness, but also it has good robustness, the algorithm is good in common uses, and it can adapt to the environment wide. Especially that the aliasing noise is in the low and high frequency band, noisy speech endpoint detection is very good, it is worthy of further improvement in the case of noise aliasing in the voice band.

References

[1] Raj B, Singh R. Classifier-based non-linear projection for adaptive endpointing of continuous speech. *Computer Speech and Language.* 2003; 17: 5-26.

[2] Tanyer SG, Ozer H. Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*. 2000; 8(4): 478-482.

[3] Karray L, Martin A. Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication*. 2003; 40: 261-276.

[4] Kuroiwa S, Naito M, Yamamoto S, et al. Robust speech detection method for telephone speech recognition system. *Speech Communication.* 1999; 27: 135-148.

[5] Ramirez J, Segura JC, Benitez C, et al. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication.* 2004; 42: 271-287.

[6] Ramirze J, Segura JC, Benitez C, et al. An efective subband OSF-based VAD with noise reduction for robust speech recognition. *IEEE Transactions on Speech and Audio Processing.* 2005; 13(6): 1119-1129.

[7] Nemer E, Goubran R, Mahmoud S. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*. 2001; 9(3): 217-231.

[8] Shen J, Hung J, Lee L. *Robust entropy-based endpoint detection for speech recognition in noisy environments*. Proc of International Conference on Spoken Language Processing, Sydney, Australia. 1998: 232-238.

[9] Ephraim Y, van Trees H LA signal subspace approach for speech enhancement. *IEEE Trans on Speech Audio Processing*. 1995; 3(4): 251-266.

[10] Klein M, Kabal P. Signal subspace speech enhancement with perceptual post filtering. *IEEE-ICASSP*. Orlando, Florida, USA. 2002: 537-540.

[11] Mittal U, Phamdo N. Signal/noise KLT based approach for enhancing speech degraded by colored noise. IEEE Trans on Speech Audio Processing. 2000; 8: 159-167.

[12] Yi H, Loizou P. CA generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans on Speech and Audio Processing. 2003; 11(4).

[13] Spib noise data [EB / OL] [2011-10-20] .http:. //spib.rice.edu/spib/select_noise.html.

[14] YAN Run-qiang, ZHU Yi-sheng. Speech endpoint detection based on recurrence rate analysis. Journal of Communications. 2007; 28(1): 35-39.doi:10.3321/j.issn:1000-436X.2007.01.006

[15] Marzinzik M, Kollmeier B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans on Speech and Audio Processing. 2002; 10: 109-118.

[16] LI Jin, WANG Jing-fang, GAO Jin-ding. Speech endpoint detection algorithm based on EMD and RP. *Computer Engineering and Applications.* 2010; 46(34): 132-135. doi:10.3778/j.issn.1002-8331.2010.34.040

[17] WANG Jingfang. Real-time voice activity robust detection. *Computer Engineering and Applications.* 2011; 47(20): 147-149. doi:10.3778/j.issn.1002-8331.2011.20.042