Advanced Engineering Forum *Online: 2011-12-22*
ISSN: 2234-991X, Vols. 2-3, pp 135-139
doi:10.4028/www.scientific.net/AEF.2-3.135

Speech Endpoint Detection in Noisy Environment Based on the Ensemble Empirical Mode Decomposition

Jingjiao Li, Dong An, Jiao Wang, Chaoqun Rong

School of Information Science & Engineering Northeastern University, China.

E-mail: 249350656@qq.com

Keywords：EEMD; ICA; Speech Endpoint Detection

Abstract：Speech endpoint detection is one of the key problems in the practical application of speech recognition system. In this paper, speech signal contained chirp is decomposed into several intrinsic mode function (IMF) with the method of ensemble empirical mode decomposition (EEMD). At the same time, it eliminates the modal mix superposition phenomenon which usually comes out in processing speech signal with the algorithm of empirical mode decomposition (EMD). After that, selects IMFs contained major noise through the adaptive algorithm. Finally, the IMFs and speech signal contained chirp are input into the independent component analysis (ICA) and pure voice signal is separated out. The accuracy of speech endpoint detection can be improved in this way. The result shows that the new speech endpoint detection method proposed above is effective, and has strong anti-noises ability, especially suitable for the speech endpoint detection in low SNR.

Introduction

The speech endpoint detection has great significance in speech signal processing. Accurate speech endpoint detection can not only improve the accuracy of speech recognition, but also reduce the quantity of computational data. However, for the existence of background noise, especially in the case of strong background noise, the accuracy of speech endpoint detection has been great affected. There are a great number of speech endpoint detection methods, such as Short-time Energy, Short-time Zero-crossing Rate, Information Entropy, Mel-Frequency Cepstrum Coefficient (MFCC), Hidden Markov Models (HMM), Wavelet Transform technology. However, these methods still have some defects, especially in low signal-to-noise ratio (SNR) conditions. Mr. Liu provided an improved sub-band adaptive entropy method [1]. The accuracy rate of detection is 69.24% at 0dB SNR conditions. Miss Ma provided a method of time-frequency variance summation [2]. The accuracy rates were 57% and 46% respectively at 0dB and -5dB conditions. Mr. JUANG provided the WE-based SRNFN algorithm [3]. The detection rate was 82.74% at 5dB.

Since the above methods can not detect speech signals accurately at low SNR, in this paper, we provide a method of endpoint detection which based on ensemble empirical mode decomposition (EEMD)[4]. EEMD is a time-frequency analysis method which does not need any prior knowledge. Its decomposition depends on the signal itself. The decomposition of the data not only has real physical meanings, but also has a higher time-frequency resolution. Therefore, this analysis method will be a great breakthrough in analyzing non-stationary and nonlinear speech signal.

Methods

The core of EEMD is EMD which decompose the signal into a number of IMFs. An IMF must satisfy the following two conditions:

1) In the whole signal set, the number of extremes and zero crossings must be either the same or at most differ by one;

2) At any point, the mean value of the envelope defined by the maxima and minima is zero.

The process of EMD decomposition is as follows:

1) Identify all the maxima of the whole signal, and then fit on the upper envelope of the signal by using cubic spline curve maxima points of interpolation;

2) Repeat the above methods, but find all the minima, fit on the lower envelope. The mean value function of the upper and lower envelope is defined as m_1, and then the first signal component can be calculated as:

$$h_1 = x(t) - m_1 \tag{1}$$

where $x(t)$ is the original speech signal.

Ideally, h_1 should be an IMF. However, it is difficult to obtain the theoretical upper and lower in reality, so cubic spline fitting is used for approximating. In the following sifting process, h_1 is defined as original signal, and by repeating the above steps we can get more satisfactory results.

$$h_{11} = h_1 - m_{11} \tag{2}$$

Repeat the sifting process k times, until the h_{1k} meets the IMF conditions.

$$h_{1(k-1)} - m_{1k} = h_{1k} \tag{3}$$

In order to ensure that the IMF components retain enough physical sense of both amplitude and frequency modulations, we have to limit the number of iteration of sifting. In the implementation of the algorithm, this can be accomplished by the following standard deviation (Standard Deviation, SD) as:

$$SD = \sum_{t=0}^{r} \left[\frac{\left| h_{1(k-1)}(t) - h_{1k}(t) \right|^2}{h_{1(k-1)}^2(t)} \right] \tag{4}$$

If the decomposing process is correct, SD should be between 0.2-0.3.

For the sake of convenience, the first IMF is defined as

$$c_1 = h_{1k} \tag{5}$$

Separate c_1 from the rest of the signal by:

$$x(t) - c_1 = r_1 \tag{6}$$

Since the residual r_1 still contains information of longer period components, we will treat it as a new data and carry out the same sifting process as described above:

$$r_1 - c_2 = r_2, \cdots r_{n-1} - c_n = r_n \tag{7}$$

When the component c_n or the residue r_n becomes so small that it is less than the small predetermined value, or when the residue r_n becomes a monotonous curve from which no more IMF can be extracted, stop the sifting process. Finally we can obtain:

$$X(t) = \sum_{i=1}^{n} c_i + r_n \tag{8}$$

When a band of the signal is discontinuous in the component, the normal decomposing of EMD will be influenced, along with the appearance of modal confusing phenomenon. The principle of the EEMD is as follows[4]:

1) A collection of white noise cancels each other out in a time-frequency ensemble mean; therefore, only the signal can continue to exist and remain in the final noise-added ensemble mean.

2) White noise of finite amplitude necessarily compels the ensemble to discover all possible solutions. The white noise makes the different scale signals reside in the corresponding IMFs, controlled by dyadic filter banks, and renders the results of ensemble mean more meaningful.

3) The decomposition result with truly physical meaning of the EMD is not the one without noise; it is assigned to be the ensemble mean of a large number trials comprising the noise-added signal.

Speech Endpoint Detection

The main reason influencing of speech endpoint detection accuracy is the composition of the noise of speech signal. The spectral entropy method can get very good detection effect in higher SNR cases, but detection accuracy will drop rapidly in low SNR cases. So, the noise speech denoising is

particularly important. In this paper, the speech signal is decomposed according to the EEMD method with the ensemble number of 40 and the amplitude of 0.2 time standard deviation of the signal. Noise usually appears in some IMF. Select "major noise" the IMF procedure is as follows.

1) The noisy speech signal is decomposed into I components by EEMD. $e(i)$ stands for the average value of the 20 frames signal energy before the ith IMF component, $E(i)$ stands for the peak of the whole signal frame energy of the ith IMF component. $a(i)$ is the weight coefficient of the Ith IMF rank defined as follows:

$$a(i) = \frac{E(i) - e(i)}{E(i)} \tag{9}$$

2) If the Ith rank IMF is "major noise", $a(i)$ will be close to a bare minimum value. If the Ith rank of IMF is "major signal", $a(i)$ will be close to 1. In this way, it can get the weight coefficients of the IMF quickly. Then IMF corresponding to the minimum $a(i)$ is treated as "major noise". Finally the "major noise" and the speech signal with noise are isolated by the ICA algorithm[5], resulting in voice component and noise component. Following is the flowchart.

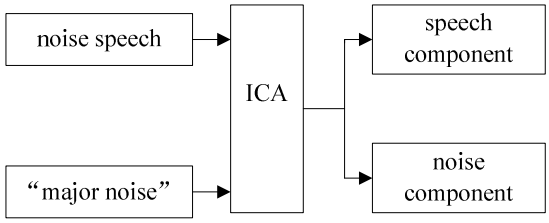

Fig. 1. Processing flow

The speech endpoints can be detected by judging the energy singularity point in noisy speech signal. Concrete steps are as follows.

1) The fragment which is chosen at the beginning of the noisy speech signal is used to estimate background noise of the speech signal. We choose the first 10th frames as noisy fragment to process. Make sure that the mean and variance energy of first 10th frames.

$$E(A_x) = \frac{1}{10}\sum_{i=1}^{10} A_x(l) \tag{10}$$

$$D(A_x) = \frac{1}{10}\sum_{i=1}^{10} [A_x(l) - E(A_x)]^2 \tag{11}$$

2) The threshold T is used for speech endpoint detection, as follow.

$$T = E(A_x) + \beta \bullet D(A_x) \tag{12}$$

where β is the constant and the value is 2 based on a great of experiments.

3) Speech segment and non-speech segment are judged by threshold T on the energy spectrum. The endpoints are signed in the speech signal.

Experiment

In the experiments, this paper we chose ten speech signals from the TIMIT (five boys and five girls). Chose noise signal from the noisex – 92, including Vehicle interior noise, F-16 noise and Factory noise. Added those noises into the pure speech according to the signal-to-noise ratio (SNR) 0dB and -5dB. In order to check the effect of endpoint detection, the speech endpoint detection method based on the spectral entropy method was defined as compared method.

In order to prove the above methods validity, selected a period of girl's in TIMIT as pure speech randomly, the content was: "She had your dark suit in greasy wash water all year". Figure 2 showed the endpoint detection results of spectral entropy method on pure speech, and we can conclude that spectral entropy method can detect speech endpoint very well.

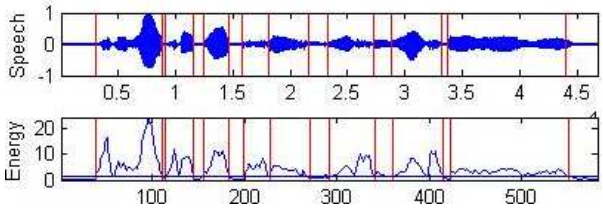

Fig. 2. The endpoint detection results of Spectral entropy method on pure speech

Figure 3 was the condition added Vehicle interior noise below 0dB SNR. Compared with the condition of pure speech endpoint detection, the spectral entropy method appeared several mistaken examining endpoints.

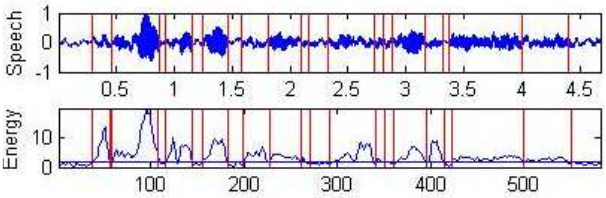

Fig. 3. SNR=0dB (Spectral entropy method)

Figure 4 was the result of speech endpoint detection using EEMD under 0dB SNR condition. Contrasted with the result of pure speech endpoint detection under the condition of 0dB, the method proposed in this paper only appeared several error detections.

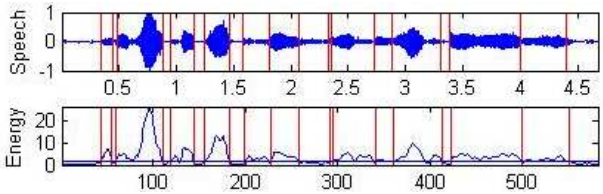

Fig. 4. SNR=0dB (EEMD)

Figure 5 showed the results of spectral entropy method in the condition of SNR below -5dB. Contrasted with the results of comparative pure speech endpoint detection, spectral entropy method under the -5dB condition appeared a large number of mistaken examining endpoints, which means failure.

Figure 6 showed the results of speech endpoint detection using EEMD in -5dB condition. Contrasted with the condition of comparative pure speech endpoint detection results, the EEMD method also appeared some examining mistakes, but detection results were far better than that of spectral entropy method, still meeting the practical needs.

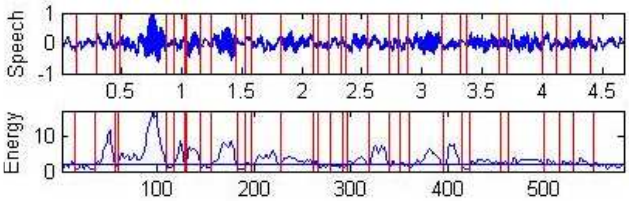

Fig. 5. SNR=-5dB (Spectral entropy method)

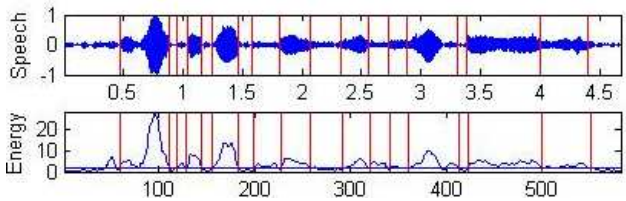

Fig. 6. SNR=-5dB (EEMD)

Figure 7 to 8 were the results of 10 sections of speech spectral entropy method and the EEMD method in this paper under the condition of 0dB and -5dB.

We used detection rate (DR) and error detection rate (EDR) to evaluate experiments result.

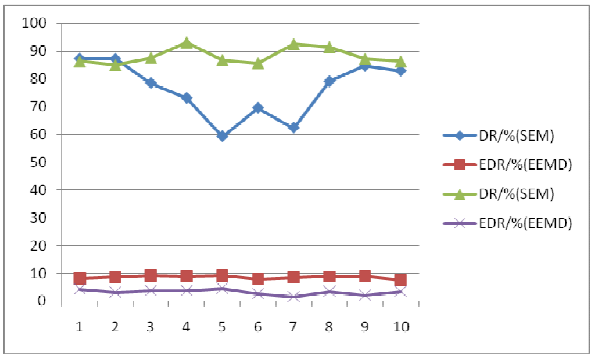

Fig. 7. SNR=0dB

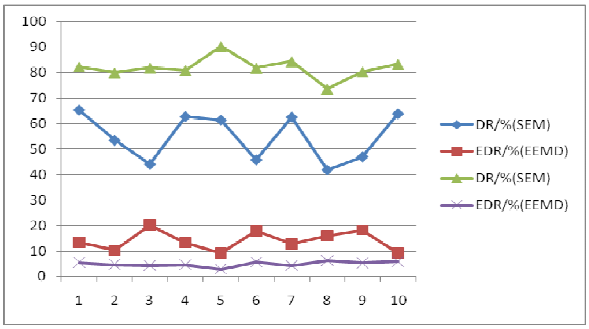

Fig. 8. SNR=-5dB

As is shown in those figures, the spectral entropy method failed in low SNR cases, while the EEMD method could still find out an appropriate adaptive threshold from the energy distribution after removing the noise, and detected the endpoint of Speech signal under the -5dB situation. Although there still existed a few undetections and mistaken examinings, but far better than spectral entropy method.

Conclusion

In this paper, a novel algorithm for speech endpoint detection based on EEMD is proposed. The noisy speech signal can be adaptively decomposed into finite IMFs by the character that the speech signal is unstable. Then, find the “major noise” self-adaptingly .After eliminating noise by the independent component analysis algorithm, detect the noisy speech endpoints at last, based on the self-adapting threshold of energy distribution. The experimental result shows that this method can detect speech signal more efficiently, while it has better stability and adaptability, especially at low SNR, such as 0dB and -5dB.

Acknowledgment

The project is sponsored by Natural Science Foundation of China (No. 60970157) and the Supported by Doctoral Fund of Liaoning province (No.20081019).

References

[1] Liu HuaPing, Li Xin, Speech endpoint detection based on improved adaptive band-partitioning spectral entropy, J. System Simulation.(2008) 51-59.

[2] MA Jingxia. Research on Noisy Voice Activity Detection Method, D. Yanshan University. 2007

[3] Juang ChiaFeng, Cheng ChunNan, Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network, J. Expert Systems with Applications. (2009) 321-332.

[4] Zhao HuaWu，N E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, J. Advances in adaptive data analysis. (2009) 1–41.

[5] Hyvarinen A，OJA E, Independent component analysis: algorithms and applications, J. Neural Networks，(2000) 411-430.