Speech Endpoint Detection Based on EMD and Higher Order Statistics in Noisy Environments

Dexiang Zhang [1, 2, a], Jiaxing Li [2, b] and Zihong Chen [2, c]

[1] Key Lab. of Intelligent Computing and Signal Processing, Anhui University, Hefei 230039, China

[2] The School of Electrical Engineering and Automation of Anhui University, Hefei, 230601, China

[a] zdxdzxy@126.com, [b] ahu_lijiaxing@126.com, [c] chenzihong315@126.com

Keywords: Endpoint Detection, EMD, Higher Order Statistics, Noisy Environments

Abstract. Accurate endpoint detection is crucial for speech recognition accuracy. This paper presents a new technique for speech endpoint detection in a noisy environment based on the empirical mode decomposition (EMD) algorithm and higher order statistics. With the EMD, the noise speech signals can be decomposed into a sum of the band-limited function called intrinsic mode functions (IMFs), which is a zero-mean AM-FM component. Then higher order statistics of the IMF components can be used to extract the desired feature for endpoint detection. In order to show the effectiveness of the proposed method, we present examples showing that the new measure is more effective than traditional measures. The experimental results show that the performance of the proposed algorithm is noticeable in the real speech signal tests with different SNR.

Introduction

Reliable endpoint detection of speech signal is a crucial preprocessing step in many speech recognition technologies and is essential in most analysis and synthesis system. The goal of speech endpoint detection is to identify the important voice of an audio segment for further processing, such as speech recognition, speech coding and communication [1].

Many algorithms have been reported for solving the speech detection problem. In early speech-detection algorithms, average zero-crossings rate (ZCR) and energy features are two algorithms, which have been the most widely used for the speech endpoint detection. However, the performance of endpoint detection by ZCR-based algorithms is not very satisfactory under noisy environments. Short time energy of the speech signal method is the most widely applied method in various endpoint detection methods. It is computationally efficient and performs adequately in a clean environment but the performance of endpoint detection by energy-based algorithms is not very satisfactory under noisy environments [2].

Higher order statistics have shown promising results in a number of signal processing applications. The Higher order statistics based signal processing method handles colored Gaussian measurement noise automatically. By extracting the feature of statistical properties for every IMF component, the proposed method can get rid of the limitation of the cumulant based methods, when the background noise is Gaussian noise. Our experiments show that the proposed method can work well for both Gaussian and non-Gaussian noise.

In this paper, we use a new technique called the empirical mode decomposition (EMD) has recently been introduced by Norden E. Huang et al. in 1998. This technique adaptively decomposes a non-stationary signal into sums of zero-mean amplitude modulation frequency modulation components. This method has quite good characters to analyze non-stationary signal and nonlinear signal. In most of the existing algorithms the Fourier transform or wavelet transform are used in signal decomposition. The speech decomposition is performed by fitting some predefined bases without satisfying its non-stationary nature. The idea behind EMD method is to decompose non-stationary signal into a number of basis function termed intrinsic mode functions. The speech signal is decomposed using EMD into the data adaptive bases up to the level of fundamental oscillations [3].

In this paper, a new endpoint detection method based on the empirical mode decomposition (EMD) algorithm and higher order statistics is proposed to accurately locate the boundaries of speech activity embedded in noisy environments.

Empirical Mode Decomposition

The EMD methods will generate a collection of intrinsic mode functions [4]. Given a signal $x(t)$, identify local maxima and minima, Then generate the upper envelope $x_u(t)$ and the lower envelops $x_l(t)$ by connecting the maxima and minima separately with cubic spline interpolation. Then the mean of the two envelops is denoted as:

$$m_1(t) = (x_u(t) + x_l(t))/2 \tag{1}$$

Thus, the first IMF $h_1(t)$ is obtained as:

$$h_1(t) = x(t) - m_1(t) \tag{2}$$

We can repeat this sifting procedure k times, until h_{1k} is an IMF, that is

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

Then, designate h_{1k} as c_1

$$c_1 = h_{1k} \tag{4}$$

where c_1 is the first IMF of the original signal. We can separate $c_1(t)$ from the rest of the data by:

$$r_1(t) = x(t) - c_1(t) \tag{5}$$

Note that the residue $r_1(t)$ still contains some useful information. We can therefore treat the residue as a new signal and apply the above procedure to obtain:

$$r_N(t) = r_{N-1}(t) - c_N(t) \tag{6}$$

The sifting process will be continued until the final residue $r_N(t)$ is a constant, a monotonic function, or a function with only one maxima and one minima from which no more IMF can be derived. Combining the equations in (5) and (6) yields the EMD of the original signal:

$$x(t) = \sum_{i=1}^{N} c_i(t) + r_N(t) \tag{7}$$

The result of the EMD produces N IMF and a residue signal. It is observed that higher order IMFs contain lower frequency oscillations than that of lower order IMFs. If we interpret the EMD as a time-scale analysis method, each IMF reflects the characteristic of each scale, which shows the intrinsic mode characteristic of non-stationary and nonlinear signal.

Higher Order Statistics Based Endpoint Detection

Cumulants and moments are higher-order covariance functions with properties that make them very useful for describing both stochastic and deterministic signals. With the EMD, the noise signals can be decomposed into different numbers of IMFs. Then, the fourth-order cumulant (FOC) can be used to extract the desired feature of statistical properties for IMF components. Since the higher-order cumulants are blind for Gaussian signals, the proposed method is especially effective regarding the problem of speech-stream detection, where the speech signal is distorted by Gaussian noise. Besides that, with the self-adaptive decomposition by the EMD, the proposed method can also work well for non-Gaussian noise [5].

If $x(n)$ is zero mean random discrete-time signal and its moments up to order *k* exist, then its *kth* order cumulant can be calculated according to their *kth* order moments. Fourth order cumulant of the random signal $x(n)$ can use type estimation:

$$C_{4x}(0) = E\{x^4(n)\} - 3(E\{x^2(n)\})^2 \tag{8}$$

When estimating higher-order statistics from finite data records, the variance of the estimators can be reduced by normalizing the input data to have a unity variance before computing the estimators. Equivalently, the fourth-order statistic may be normalized by the second power of the data variance, thus the normalized kurtosis is

$$C_{4x}(0) = \frac{E\{x^4(n)\}}{(E\{x^2(n)\})^2} - 3.0 \tag{9}$$

In the process of endpoint detection, the sum of the higher-order statistics values over duration of frames is first evaluated. Some thresholds are then used to detect the beginning and ending boundaries of the embedded speech segments in a continuous utterance.

A speech signal is first decomposed into often finite IMFs by the EMD, as shown in Equation (7). The normalized kurtosis of the j^{th} IMF in terms of the definition of Equation (9) is:

$$C_j(0) = \frac{E\{\mathrm{IMF}_j^4(n)\}}{(E\{\mathrm{IMF}_j^2(n)\})^2} - 3.0 \tag{10}$$

During the decomposition of the EMD, on each little period of time, IMFs with the minimal scale are obtained first, then are IMFs with larger scales, in the end is the IMF with the maximal scale. For a noisy speech signal, the resulting IMFs can have different Gaussian characteristics, which help to discriminate the speech sounds from noise. On the other hand, the larger scales have very low amplitudes, which, according to the FOCs, are very small compared to the other IMFs, and thus it is not necessary to calculate these posterior IMFs. This helps to reduce the computing time.

Experiments and Results

The first experiment signal is a pure speech signal; the sampling frequency is set to 10 kHz with 16-bit amplitude resolution. The speech signal is segmented into frames of length 25.6ms with 12.8ms overlap. Each frame is labeled manually as 0.5 for voiced frame and 0 for unvoiced one [6]. The experimental results are shown in Fig.1. The experiment results in the fig. 1 are shown that the proposed method can accurately extract the speech signal.

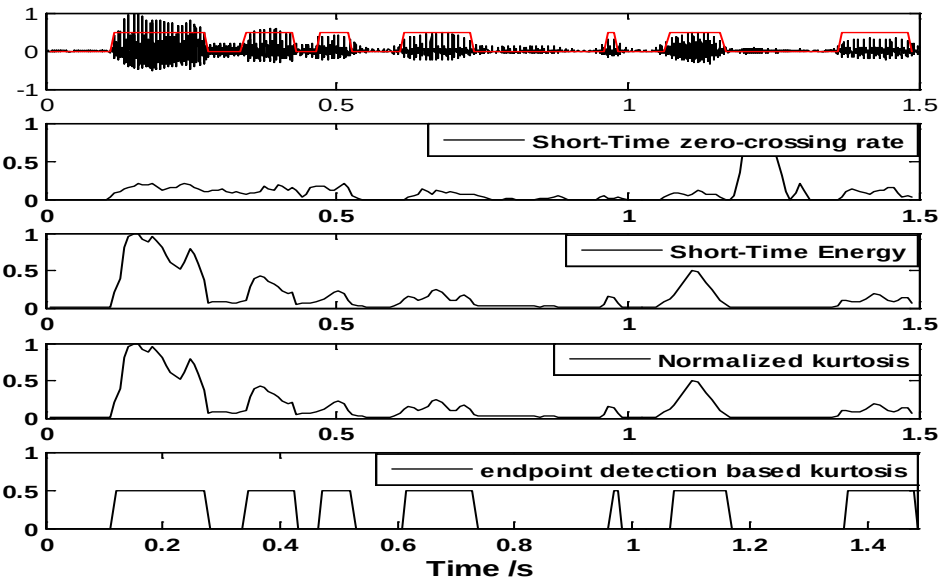

Fig. 1 Results of the endpoint detection using EMD and normalized kurtosis

The second experiment data is the pure speech signal in Figure 1 corrupted by White noise and SNR=0dB. The experimental results are shown in Fig. 2.

The test speech signal contains 108 frames. The spectral entropy of first ten IMFs is computed in terms of the definition of Equation (10). Moreover, the total signal normalized kurtosis as a function

of time is the sum of the energies of the first nine IMFs normalized kurtosis. The experiment results in the fig. 2 are shown that the proposed method can accurately extract the speech signal.

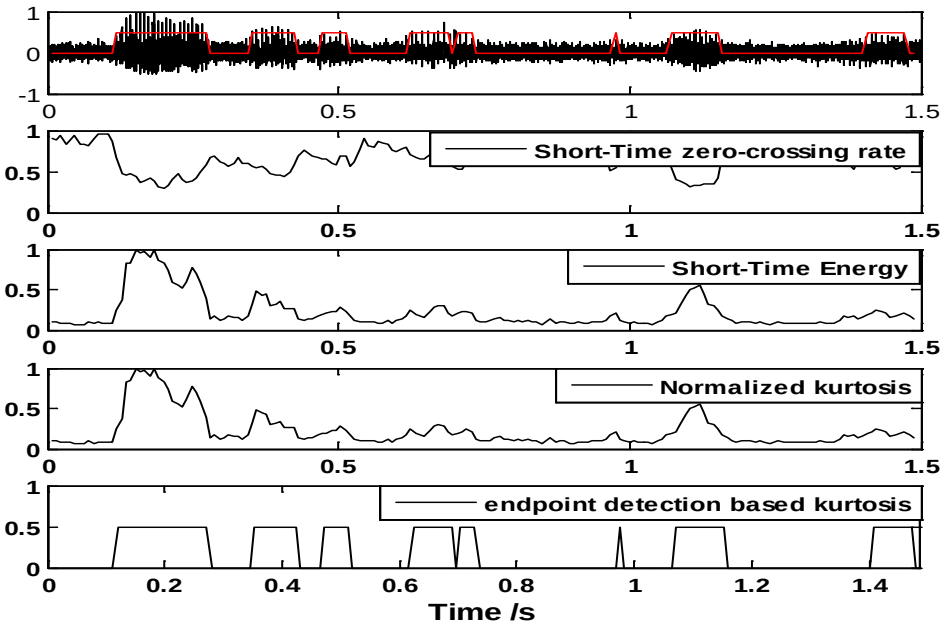

Fig. 2 Results of the endpoint detection in noisy environment using EMD and normalized kurtosis

Conclusion

In this paper, a method based on the empirical mode decomposition (EMD) algorithm and normalized kurtosis is proposed to detect endpoints of a speech signal embedded in noise. The EMD is well adapted to non-stationary signals and has an excellent time resolution. The EMD and normalized kurtosis gives good endpoints detection of a speech signal corrupted with noise. Processing of a large number of signals and comparisons to exiting methods such as short time energy and short time zero crossing rates are necessary to show the robustness and the effectiveness of this method. Our experiments show the proposed method can accurately extract the speech signals in differently background noises.

Acknowledgment

The support from the Chinese National Science Foundation Grant (No.61272025) for this research is gratefully acknowledged.

References

[1] G. Tanyer, H. Ozer, IEEE Trans. Speech Audio Processing, Vol. 8 (2000), p. 478

[2] Zhang Yong, Chen Bin. Journal of University of Electronic Science and Technology of China, Vol. 36(2007), p. 8

[3] N. E. Huang, Z. Shen and S. R. Long, Proceedings of the Royal Society London A, Vol. 454 (1998), p. 903

[4] I. Djurovic and L.J. Stankovic,IEEE Transaction on Signal Processing, Vol. 49 (2001), p. 2985

[5] Wang Rangding, Chai Peiqi. Information and Control (In Chinese), Vol. 33 (2004), p. 77

[6] De-Xiang Zhang, Xiao-Pei Wu, and Zhao Lv. Journal of Electronic Science and Technology, Vol.8 (2010), p. 183