# A Novel Algorithm to Robust Speech Endpoint Detection in Noisy Environments

Li Yi*#, Fan Yingle*
*Institute of Biomedical Engineering and Instrument
Hangzhou Dianzi University
Hangzhou, Zhejiang Province, 310027 China
#School of Biomedical Engineering &Instrumentation
Zhejiang University
Hangzhou ,Zhejiang 310017 P.R. China

***Abstract* -Accurate endpoint detection is crucial for good speech recognition accuracy. A new algorithm based on C0 complexity measure is proposed in this paper. In the comparison of the new algorithm to the other traditional methods such as energy, ZCR, spectral-entropy etc., for the continuous speech and isolate word speech, the C0 complexity feature proved effective. The proposed algorithm is shown to be well suited for the detection of speech endpoint and is very robust for different types of noise, especially for low SNR**

## I. INTRODUCTION

Endpoint detection, which aims at distinguishing speech and non-speech segments using signal processing and pattern recognition, is considered as one of the key preprocessing components in automatic speech recognition (ASR) systems.There have many attempts to solve the endpoint detection problems over the past several decades. Computing the energy of speech signal is a computationally simple operation compared to extracting other features, such as LPC derived cepstrum coefficients (LPCC), mel-frequency cepstrum coefficients (MFCC) and so on, which have been found to work well but are time and computationally intensive [1]. As for energy-based methods, most of the algorithms are based on simple parameters such as energy contours and zero crossings [2]. Sometimes several features are combined to detect speech endpoint. Most of these methods have been shown to be effective for endpoint detection. However, sometimes, they still fail, especially in a high noisy environment.

Most of these methods such as linear prediction are substantially based on the basic hypothesis, which is the nonlinear system can be described approximately by the linear system when the segment is small enough. The analysis method can be understood easily in theory with simple computing and they are the focus of the research.

With the development of study, the disadvantage of the subsection linear analysis method cause the researchers notice, for example, the performance of the speech recognition, speaker identification, speech synthesize and speech code system can't be improved further. Therefore, researchers pay more and more attention to the nonlinear signal analysis method. Aerodynamics indicates that the speech signal is nonlinear, the chaos characteristics of speech signal has been proved [3]. We address this problem from the point of view of fractals and chaos. A new nonlinear endpoint detection method is proposed, which based on the $C_0$ complexity measure feature, one new nonlinear feature, which can reflect the essentially nonlinear characteristics of speech signal. The comparison of the proposed method to the spectral entropy has proved the proposed method to be well suited for the speech endpoint detection [4].

## II. TRADITIONAL METHOD OF SPEECH ENDPOINT DETECTION

Endpoint detection has been studied for decades and many algorithms have been proposed. Most of these methods have the followed problems.(1)All features used in speech endpoint detections is linear feature, the nonlinear features of speech are often ignored. (2)Most methods perform well in quiet environments, but degrade rapidly in noise environments.

### *A. Short-time Energy*

Among various endpoint detection approaches, energy-based methods[5] are the most widely applied solution to this problem. The idea of using energy to detect endpoints is that, in a clean environment, the energy (or power) of a voiced segment is higher than a non-voiced one in a stream of speaker's utterance. In these methods, a fixed-length window is defined first. It is then floated on the input utterance. By continuously monitoring the utterance through the window, the starting point can be found once when the short-time energy of the window is higher than some beginning threshold. Similarly, the ending point too can be located when the short-time energy of the window is lower than some ending threshold. However, the performance of endpoint detection by energy-based algorithms is not very satisfactory under noisy environments. It's difficult to differentiate the desired voice and unexpected background noise, such as the sound from opening or closing a door, cough sound, shaking sound from engine and so on.



1-4244-0737-0/07/$20.00 ©2007 IEEE

*B. Zero Crossing*

Another well-known measure is the zero crossing [6] that counts the number of times the sampled sequence changes signs. The definition of zero crossing is as followed.

$$Z_n = \frac{1}{2}\sum_{n=0}^{N-1}\left|\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]\right| \quad (1)$$

Where sgn[x (n)] is symbol function, defined as

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

But this feature is too unstable to be used on endpoint detection. It is often accompanied with energy. But the help of zero crossing is sometimes hard to be seen.

*C. Energy -Zero Crossing Product*

This algorithm is based on energy and zero-crossing parameters and a set of decision rules and threshold settings. It is widely used in many speech applications. The definition of Energy-Zero Crossing Product is that short-time energy multiplies Zero Crossing parameter [7]. This algorithm is fast and very practical: the speech signal is acquired at the same time as the word boundary detection is done.

*D. Speech Endpoint Detection Based on Cepstrum Feature*

Cepstrum feature is a good characteristic parameter of speech signal, one description method of speech signal in frequency domain. Cepstrum can be obtained by FFT transform of signal and inverse FFT transform after amplification changed to the Log function of itself. The Parsavel theorem proved the cepstrum distance can indicate the distance of the signal frequency spectrum, thus cepstrum distance can judge the speech endpoint [8].

## III. THE ALGORITHM DESCRIPTION

*A .Complexity Measure Theory*

The nonlinear analysis method for the nonlinear time series mainly include Lyapunov exponent, fractal dimension, entropy measure and several computation complexity measures. In recent years, complexity measures especially cause researcher notice, which have been applied in many fields, such as EEG signal analysis. These applications have shown that the complexity measure is a useful tool for signal processing [9].

The concept of complexity and its importance was raised by von Neumann in the 1950s'. In 1965, Kolmogorov gave the concrete description of definition of complexity, now known as the algorithmic complexity. The work of Lempel and Ziv and Kaspar and Schuster gave the procedures for computing complexity [10]. Their method can obtain the measure of information, which is called Kolmogorov Complexity. With the development of the practical symbol dynamics in recent years, many other complexity measures are given such as structure complexity measures, dynamics complexity measure, $C_1/C_2$ complexity measure. These methods, in brief, start by turning the time series to symbol sequence, which is called the coarse granulation process, then using the complexity of symbol sequence denotes the complexity of the time series. The coarse granulation process leads to much useful detail information lost. When time series include some component with low frequency and high amplification, the result of this operation becomes too simple. This case is called excessive granulation. The $C_0$ complexity measure applied in this paper can describe the nonlinear characteristic of signal and avoid t excessive granulation.

*B. $C_0$ Complexity measure*

Complexity Movement is commonly composed of orderly movement and stochastic movement. The part of stochastic movement is the basis of $C_0$ complexity measure described. Assume there is a time series of complexity movement $x(t)$, which includes the time series of orderly movement and stochastic movement. Assume the time series of orderly part is $x_1(t)$, the relation function between $x_1(t)$ and $x(t)$ denoted as $f(x)$, which is shown as followed.

$$x_1(t) = f[x(t)] \quad (2)$$

Subtract x1(t) from x(t), the subtract output is the stochastic movement. Describe it simple with transform g(x).

$$A_0 = g[x(t)] \quad (3)$$

$$A_1 = g[x_1(t) - x(t)] \quad (4)$$

$A_0$ indicates the measurement of the whole time series of complexity movement, while $A_1$ indicates the part of stochastic movement. Then, the definition of $C_0$ Complexity measure is given by:

$$C_0 = \lim_{t\to\infty}\frac{A_1}{A_0} \quad (5)$$

For the above formula, when $x_1(t)$ is the main part of $x(t)$, $C_0$ tends to be 0. This indicates the dynamic conduct of the system is completely orderly without stochastic component. On the contrary, when $x_1(t)$ occupies the small part and stochastic movement occupies the much bigger part, $C_0$ tends to be 1, which indicates the system is completely stochastic. Then the increase of $C_0$ complexity measure accompanies with the increase of the stochastic component.

$C_0$ Complexity measure is applied to speech signal analysis, includes the followed steps, where $x(t)$ is the speech signal after preprocessing such as enhancement, framing, windowing and so on.

*Step* 1.The FFT transform of $x(t)$ is described as.

$$x(k) = F[x(t)] \quad (6)$$

*Step* 2.The mean value of amplification spectrum $\bar{x}$ obtained.

$$\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x(k), \mathit{Where}\ 1 \leq k < N \quad (7)$$

Where $k$ is variable in frequency domain, $N$ is the length of $x(k)$ , which indicates the maximum of $k$. In practical operation, $\bar{x}$ can multiply a coefficient $a$, ($a$ >=1, constant), thus the orderly part can be regulated. The higher frequency component can be deemed to the offer of the orderly part, while the lower and equal part is the offer of stochastic part. Here, we only get the offer of orderly part.

$$x'(k) = \begin{cases} x(k), & if \quad x(k) > \bar{x} \\ 0, & if \quad x(k) <= \bar{x} \end{cases} \tag{8}$$

*Step* 3. The frequency spectrum of the orderly part $x'(k)$ with the invert FFT transform $F^{-1}(.)$, $x_1(t)$ can be gotten by:

$$x_1(t) = F^{-1}[x'(k)] \tag{9}$$

Here, the time series of orderly part $x_1(t)$ is found out.

*Step* 4. Using the following formula,

$$A_0 = \int_0^\infty |x(t)| dt \tag{10}$$

$$A_1 = \int_0^\infty |x(t) - x_1(t)| dt \tag{11}$$

With the formula (4), the $C_0$ complexity measure is obtained.

## IV. Experiment Evaluation

### A. Database

The databases used in the experiments include three speech databases, the English continuous speech database YOHO, one Chinese continuous speech database and one isolate word speech database.

The English continuous speech database for this work is known as the YOHO Corpus, which was collected by ITT under a U.S government contract. The YOHO database was the first large-scale, scientifically controlled and collected, high-quality speech database for speaker-verification testing at high confidence levels, and test plans have been developed for its use. The YOHO Corpus supports development, training and testing of speech analysis systems that use limited vocabulary, free-text input. The particular vocabulary employed in this collection consists of two-digit numbers ("thirty-four", "sixty-one", etc), spoken continuously in sets of three (e.g."26-81-57", say: twenty-six, eighty-one, and fifty-seven). YOHO was designed for U.S government evaluation in "office" environments.

The continuous speech corpus used to evaluate speech endpoint detection is composed with speech samples from 50 different speakers (include 25 male and 25 female), sampled at 44100Hz, with resolution of 16 bits per sample. Every speaker provided ten repetitions of the one sentence spoken in Chinese language. Every speech sample was about 300ms long. The isolated speech corpus is these speaker read the "A","B", etc, each time read one letter.

For generating the noisy speech files, we used different types of noise available from NOISEX 92 database. Four sorts of noise are considered in this paper: white noise, F-16 cockpit noise, factory noise and office noise. To set up the noisy speech database for testing, we added the prepared noisy signals to the recorded speech signals with different SNR including 0, 5, 10, 15, 20 and 30 dB. The speech waveform and the noise waveform are each calibrated and the noise segment is randomly selected from the noise tile.

### B. Experiment results

There are two possible ways to evaluate the correctness of an endpoint detection algorithm: one is to compare the detected results to hand labeled ones, and the other is to pass the detected words through a speech recognizer and compare the recognition rates [6]. Here, we choose the first option for the most straightforward comparison.

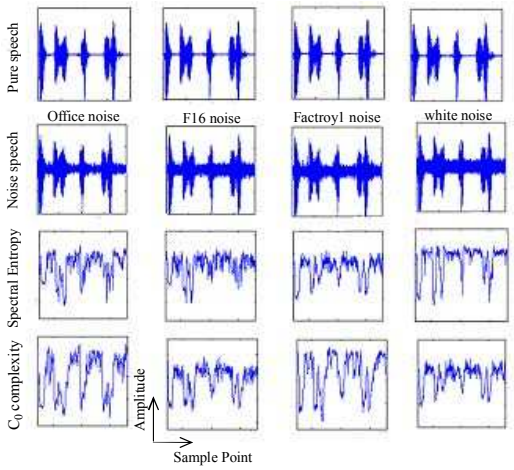


Fig.1. The $C_0$ complexity feature curves of continuous speech at low SNR (0dB)

In experiments, speech signal have windowed with hamming, the length of the window is 256 points, the overlap between the frames is 50%, the length for FFT transform is also 256 points. Fig.1 describes the $C_0$ complexity feature curves of continuous speech at low SNR (0dB). The first line in figure 1 is the speech in quiet circumstance .The second line is the noise speech with office noise, F-16 cockpit noise, factory noise and white noise. The third line is the correspondence spectral entropy feature curve. The fourth line is the $C_0$ complexity feature curve.

In this paper, we apply double adjustment value to endpoint detection. We evaluated the two algorithms for the four noise conditions at different SNR. The results are presented in Table 1. The percentage of correctness due to the $C_0$ complexity feature, spectral-entropy feature detection method is presented.

TABLE I
Detection correctness due to $C_0$ complexity feature ($C_0$) and Spectral Entropy (SE) %

| Noise | White noise | | F16 noise | | Factory noise | | Office noise | |
|---|---|---|---|---|---|---|---|---|
| | SE | $C_0$ | SE | $C_0$ | SE | $C_0$ | SE | $C_0$ |
| 30dB | 99.4 | 99.5 | 95.8 | 95.8 | 100 | 98.8 | 95.8 | 97.8 |
| 20dB | 98.2 | 98.0 | 95.0 | 95.0 | 100 | 98.0 | 95.0 | 97.2 |
| 15dB | 95.3 | 96.5 | 94.9 | 97.8 | 97.3 | 94.5 | 94.8 | 93.8 |
| 10dB | 90.5 | 92.8 | 89.0 | 90.5 | 89.5 | 91.4 | 89.5 | 88.5 |
| 5 dB | 87.6 | 89.7 | 72.0 | 79.5 | 82.6 | 87.5 | 71.7 | 82.5 |
| 0 dB | 67.4 | 75.3 | 51.6 | 79.5 | 67.2 | 80.5 | 51.7 | 77.5 |

As can be seen in Table 1, the detection correctness between the two method is approximate with the four noise at high SNR (>=10dB), while there is obvious difference between two methods at lower SNR (<=10dB). For example, compared to the spectral entropy endpoint algorithms, the $C_0$ complexity feature algorithm gives the more accurate

endpoint. The detection correctness due to $C_0$ can get 75.3% correctness of spectral entropy is just 67.4%. With the office with white noise at SNR 0dB, while the corresponding noise at SNR 0dB, the detecting correctness of $C_0$ is 77.5%, while that of spectral entropy is only 51.7%.

## V. CONCLUSION

Contrast to the other endpoint algorithm, the nonlinear algorithm with $C_0$ complexity feature has three characteristics. Firstly, the detection correctness is higher. Secondly, the robustness is well and all kinds of noise can't influence the performance of this algorithm. Finally, the algorithm computes simple and can be applied to real-time endpoint detection. In the experiment, for the continuous speech and isolate word speech, the $C_0$ complexity feature proved effective. The methods can be well suited for the robust recognition system. Based on fractals and chaos theory, we proposed a new algorithm which uses a nonlinear parameter $C_0$ complexity feature. This is an absolute new method; provide researcher a new development direction. Nevertheless, according to our results, there is still room for improvement .An important future direction will be trying to apply more complexity measures to speech endpoint detection.

## REFERENCES


[1] Yiying Zhang, Xiaoyan Zhu, Yu Hao, Yupin Luo, "A Robust and Fast Endpoint Detection Algorithm for Isolated Word Recognition," IEEE ICIPS-97,pp.1819-1822, 1997.

[2] E. Dermatas, N. Fakotakis, G. Kokkinakis, "Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment," In Proc. IEEE ICASSP-91, pp.733-736, 1991.

[3] C. Thompason, A. Mulpor, V. Mehta., "Transition to Chaos in Acoustically Driven Flow," Acoust Soc Am, pp.2097-2103,1991.

[4] Wang Rangding, Chai Peiqi. "An Improved Speech Endpoint Detection Method Based on Spectral Entropy," Information and Control (In Chinese), Vol.33, Issue 1,pp.77-81,2004.

[5] J.C. Junqua, B. Mak, and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", IEEE Trans. On Speech and Audio Processing, Vol. 2, No., 3, pp. 406-412, Apr. 1994..

[6] Sahar E. Bou-Ghazale and Khaled Assaleh, "A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition". In Proc. IEEE ICASSP - 02, Vol. 4, 2002, pp.3808–3811.

[7] Lingyun Gu and Stephen A. Zahorian, "A New Robust Algorithm for Isolated Endpoint De-tection", In Proc. lEEE ICASSP - 02, Vol. 4, ,2002, pp. 4161– 4164.

[8] W. Gin-Der and L. Chin-Teng, "Word boundary detection with mel-scale frequency bank in noisy environment", Speech and Audio Processing, IEEE Transactions on, vol. 8, 2000,pp. 541

[9] A .Lempel, J. Ziv, "On the Complexity of Finite Sequences ," IEEE Trans. on information theory, Vol.22, pp.75-88,1976.

[10] P. Janume, M. Dosal, "Robust Speech Activity Detection Using LDA Applied to FF Parameters ," Proceeding of ICASSP, pp.57-572,2005.