A New Endpoint Detection Method Based on Chaotic System Features for Digital Isolated Word Recognition System

A New Endpoint Detection Method Based on Chaotic System Features for Digital Isolated Word Recognition System

장 한, 정길도
Xian Zang, Kil To Chong

Abstract - In the research of speech recognition, locating the beginning and end of a speech utterance in a background of noise is of great importance. Since the background noise presenting to record will introduce disturbance while we just want to get the stationary parameters to represent the corresponding speech section, in particular, a major source of error in automatic recognition system of isolated words is the inaccurate detection of beginning and ending boundaries of test and reference templates, thus we must find potent method to remove the unnecessary regions of a speech signal. The conventional methods for speech endpoint detection are based on two simple time-domain measurements - short-time energy, and short-time zero-crossing rate, which couldn't guarantee the precise results if in the low signal-to-noise ratio environments. This paper proposes a novel approach that finds the Lyapunov exponent of time-domain waveform. This proposed method has no use for obtaining the frequency-domain parameters for endpoint detection process, e.g. Mel-Scale Features, which have been introduced in other paper. Comparing with the conventional methods based on short-time energy and short-time zero-crossing rate, the novel approach based on time-domain Lyapunov Exponents (LEs) is low complexity and suitable for Digital Isolated Word Recognition System..

Key Words : Digital Isolated Word Recognition; Time-domain; Time-dependent Lyapunov exponent

1. Introduction

Endpoint detection, which aims at distinguish the speech and non-speech segments from digital speech signal, is considered as a crucial part of speech signal processing, such as automatic speech recognition. A good endpoint detector can improve the accuracy and speed of a speech recognition system. In particular, a major source of error in automatic recognition system of isolated words is the inaccurate detection of beginning and ending boundaries of test and reference templates, thus it is essential to locate the regions of a speech signal that correspond to each word. Furthermore, an appropriate scheme for locating the endpoint of a speech signal can be used to eliminate significant computation by making it possible to process only the parts of the input that correspond to speech.

The conventional endpoint detection methods are mainly based on the simple energy detector, which could performs adequately for clean speech. Most of these methods use short-time energy and zero-crossing as a useful algorithm for locating the beginning and ending point under the high signal-to-noise condition, but will degrade seriously in noisy circumstance.

저자 소개
* 媛 媚 : 全北大學 電子情報工學科 碩士課程
* 丁吉道 : 全北大學 電子情報工學科 敎授 · 工博

Aerodynamic indicates that the speech signal is nonlinear, the chaos characteristic of the speech signal has been proved. We address this problem from the point of view of chaos. A novel nonlinear endpoint detection method is proposed, which is based on time-dependent Lyapunov exponents. Comparing with those algorithms, the method just carry out the calculation basing on the time-domain waveform of speech signal, therefore it's low complexity. Experimental results show a good performance in extracting the speech segments from utterance containing a variety of background noise, and also has a good performance in resisting noise.

2. Conventional Method of Speech Endpoint Detection

Endpoint detection has been studied for decades and many algorithms have been proposed. Most of these methods have the following 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 quite environment, but degrades rapidly in noise environments.

The conventional algorithm for solving the problem of endpoint detection of a speech utterance is based on two simple time-domain measurements: short-time energy, and short-time zero-crossing rate. They are combined to serve as the basis of a useful algorithm for endpoint detection in many previous research work [1].

In general, we can define the short-time energy as

$$E_m = \sum_{n=m}^{m+N-1} s_w^2(n) \tag{1}$$

here, $s_w(n)$ is the speech signal after windowing.

This expression can be written as

$$E_m = \sum_{n=m}^{m+N-1} \hat{s}^2(n) \bullet h(n-m) \tag{2}$$

where

$$h(n) = w^2(n) \tag{3}$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & else \end{cases} \tag{4}$$

where $w(n)$ is Hamming window. N is the number of points in one frame.

Another parameter is short-time zero-crossing rate. An appropriate definition is

$$Z_m = \frac{1}{2}\sum_{n=1}^{N-1} \left| sgn\left[s_w(n)\right] - sgn\left[s_w(n-1)\right] \right| \tag{5}$$

where $sgn\left[s_w(n)\right]$ is symbol function, defined as

$$sgn[x] = \begin{cases} 1, & x \ge 0 \\ -1, & x < 0 \end{cases} \tag{6}$$

Based on the combination of the two features, one such algorithm for locating the beginning and end of a speech signal was studied by Rabiner and Sambur [2] in the context of an isolated word speech recognition system [3]. This algorithm is fast and practical: the speech signal is acquired at the same time as the word boundary detection is done. However, it couldn't guarantee the success in noisy environment.

3. Time-dependent Lyapunov Exponents Algorithm

As we know, in mathematics, the Lyapunov exponent of a dynamic system is a metric that characterizes the rate of separation of infinitesimally close trajectories. Consider two points in a space, X_0 and $X_0 + \Delta x_0$, each of which will generate an orbit in that space (Fig.1). These orbits can be thought as parametric functions of a variable that is something like time. If we use one of the orbits as reference orbit, then the separation between the two orbits will also be a function of time.

For chaotic points, the function $\Delta x(X_0, t)$ will behave erratically. It's thus useful to study the mean exponential rate of the divergence of the two initially close orbits using the formula

$$\lambda = \lim_{\substack{t\to\infty \\ |\Delta x_0 \to 0|}} \frac{1}{t} ln \frac{|\Delta x(X_0, t)|}{\Delta x_0} \tag{7}$$

This number, called the Lyapunov exponent "λ", determines the predictability of a dynamic system. A positive LE is usually taken as an indication that the system is chaotic.

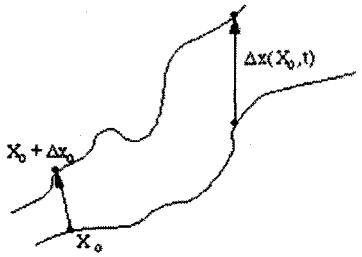


Fig.1 The separation of two orbits from the two points

Rosenstein, Collins, and De Luca(1993) proposed a method to estimate the Lyapunov exponents from a time series composed a few samples. Good results were obtained for the Lyapunov exponent estimation of known systems using lass than 1000 samples. This characteristic is very important when dealing with speech, since a speech signal can be considered stationary only during a small window of approximate 30ms(Deller et al., 1987). Furthermore, it allows the correct estimation of Lyapunov exponents from speech windows, using speech recorded at low sample rates, such as telephone speech.

We adopt the rationale of Lyapunov exponent [4-5] and make some conversion to serve for the speech recognition system. In our works, each isolated word signal was sampled at 8 kHz for 1sec, thus we got 8000 samples distributed in time-domain. The calculation of LEs is outlined as follows: the first step is to devide the time-domain waveform of utterance signal into 100 frames, each of which is 10 ms. Then after adding Hamming window, we do the following work in each frame:

a. Find the maximum and minimum amplitude during this frame;

b. Segment the amplitude region between the maximum and minimum value into many small sects based on the sample number in the frame. The sketch map was shown in Fig.2.

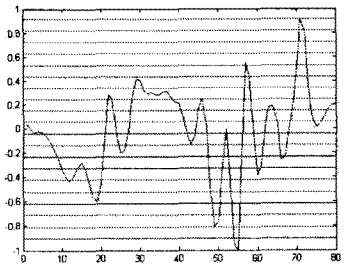


Fig. 2 The sketch map of segmentation in each frame (In fact, the number of horizontal line is equal to the number of samples in the frame)

c. In each small region, Check the sample number n, if $n \ge 2$, compute the amplitude dispersion d0 d1 d2···dn between two close samples from the first sample in the

region;

d. For the each couple of samples in step c, find another two samples which are next to the previous respectively, maybe the two samples found now are not in the region. Then compute the amplitude dispersion of the new couple of samples d0', d1', d2'…dn';

e. Compute Lyapunov exponent using the following formula:

$$\lambda = \frac{\sum_{i=0}^{n} \log_2(d'(i)/d(i))}{n+1} \qquad (8)$$

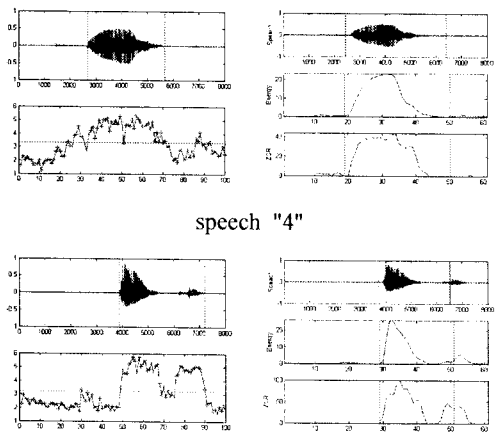
f. After finish the computation for all the small regions one by one, choose the mean of the exponents as the final Lyapunov exponent of the frame.

g. Then choose the next frame and repeat the work mentioned above.

h. Set a threshold among all of the LEs to filter the noise segments. Thus we realize the discrimination between speech and background noise.

4. Simulation Results

We use the software "GoldWave" to record English digitals 0-9 at 8 kHz sampling rate in the lab environment. The corpus contains 20 speakers, every person provided 2 repetitions from 0-9. Then we carry out the endpoint detection using the proposed method, and compare it with the conventional algorithm. To simulate speech production in noisy conditions, each speech was added noise to obtain noisy speech. Several levels of signal-to-noise (SNR) have been considered. The noise database is NOISEX92. We choose the representative noise white noise, F-16 cockpit noise, pink noise and babble noise.



speech "4"

speech "8"

Fig.3 The endpoint detection results on clean speech. The left column-Lyapunov exponents; The right column-conventional algorithm

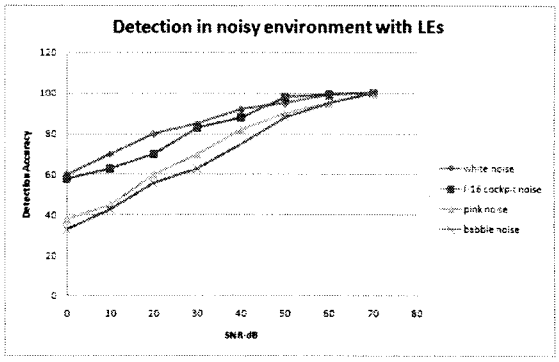


Fig.4 The detection accuracy in noisy environment. x-coordinate is the SNR(dB), y-coordinate is the detection accuracy in percentage

5. Conclusion

Endpoint detection of speech is very important for isolated word recognition. The characterization of a speech signal using non-linear dynamical features has been the focus of intense research lately. In this paper, the proposed algorithm of Lyapunov exponents provides new information to better characterize it, and perhaps point out a way of improving the accuracy of speech recognition。

Simulation results show that not only the proposed method with low complexity could extract the speech segments more accurately than conventional algorithm, but also has a good performance in decreasing SNR environments.

참 고 문 헌

[1] Zebulum, R.S.; Vellasco, M.; Perelmuter, G.; Pacheco, M.A.; " A comparison of different spectral analysis models for speech recognition using nueral networks", IEEE 39th Midwest symposium on Circuits and Systems, 1996., Volume 3,18-21 Aug. 1996 Page(s):1428 - 1431 vol.3.

[2] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., vol. 54, No. 2, pp. 297-315, February 1975.

[3] M. R. Sambur and L. R. Rabiner, "A Speaker Independent Digit-Recognition System", Bell Syst. Tech. J., vol. 54, No. 1, pp. 81-102, January 1975.

[4] Kokkinos, I.; Maragos, P., "Nonlinear peech analysis using models for chaotic systems", Speech and Audio Processing, IEEE, volume 13, Issue 6, Nov. 2005 Page(s): 1098-1109.

[5] Adriano. Petry, D. A. C. Barone, "Preliminary experiments in speaker verification using time-dependent largest Lyapunov expontent", Computer Speech and Language, 17 (2003), 403-413.