논문 2009-46SC-5-2

음성인식을 위한 혼돈시스템 특성기반의 종단탐색 기법

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

장 한[*], 정 길 도[**]

(Xian Zang and Kil To Chong)

요 약

음성 인식 연구에서 잡음이 있는 상태에서 음성 발음상의 시작점과 종단점을 찾는 것은 매우 중요하다. 기존 음성인식 시스템의 오차는 대부분 참고템플릿의 시작점과 종단점을 왜란이나 잡음으로 인해 자동적으로 찾지 못했을 경우 발생한다. 따라서 음성 신호상에서 필요 없는 부분을 제거할 수 있는 방법이 필요하다. 기존의 음성 종단점을 찾는 방법으로는 시간도메인 측정 방법, 미세시간 에너지 분석, 영교차율 방법이 있다. 위의 방법들은 저주파 신호 노이즈의 영향에 정밀성을 보장을 못한다. 따라서 본 논문에서는 시간영역상에서 리야프노프 지수를 이용한 종단점 인식 알고리즘을 제안하였다. 기존의 방법들과의 비교를 통해 제안한 방법의 성능 우수성을 보였으며, 시뮬레이션 및 실험을 통해 잡음환경에서도 음성종단 인식이 가능함을 보였다.

Abstract

In the research field of speech recognition, pinpointing the endpoints of speech utterance even with the presence of background noise is of great importance. These noise present during recording introduce disturbances which complicates matters since what we just want is to get the stationary parameters corresponding to each speech section. One major cause of error in automatic recognition of isolated words is the inaccurate detection of the beginning and end boundaries of the test and reference templates, thus the necessity to find an effective method in removing the unnecessary regions of a speech signal. The conventional methods for speech endpoint detection are based on two linear time-domain measurements: the short-time energy, and short-time zero-crossing rate. They perform well for clean speech but their precision is not guaranteed if there is noise present, since the high energy and zero-crossing rate of the noise is mistaken as a part of the speech uttered. This paper proposes a novel approach in finding an apparent threshold between noise and speech based on Lyapunov Exponents (LEs). This proposed method adopts the nonlinear features to analyze the chaos characteristics of the speech signal instead of depending on the unreliable factor-energy. The excellent performance of this approach compared with the conventional methods lies in the fact that it detects the endpoints as a nonlinearity of speech signal, which we believe is an important characteristic and has been neglected by the conventional methods. The proposed method extracts the features based only on the time-domain waveform of the speech signal illustrating its low complexity. Simulations done showed the effective performance of the proposed method in a noisy environment with an average recognition rate of up 92.85% for unspecified person.

Keywords : Digital Isolated Word Recognition; Time-domain; Time-dependent Lyapunov Exponents

Ⅰ. 서 론

The endpoint detection's aim to distinguish the speech segment from the non-speech segment of a digital speech signal is considered as a crucial part of

[*] 학생회원, 전북대학교 제어계측공학과
(Control and Instrumentation Department, Chonbuk National University)
[**] 정회원, 전북대학교 전자정보공학부
(Electronics and Information Departemtn, Chonbuk National University)

접수일자: 2009년6월26일, 수정완료일: 2009년7월17일

speech signal processes such as in an automatic speech recognition system wherein a good endpoint detector can improve the accuracy and speed of the system. Since an inaccurate detection of the beginning and ending boundaries of the test and reference templates is a major source of error in an automatic recognition system of isolated words, it is essential to locate the regions of a speech signal that correspond to each word. Furthermore, an appropriate scheme for locating the beginning and end of a speech signal can be used to eliminate significant computational tasks by making it possible to process only the parts of the digital signal input that correspond to speech.

In the last several decades, a number of endpoint detection methods have been developed and they can be categorized approximately into two classes, the threshold-based[1~3] and the pattern-matching methods[4~5]. The threshold-based method first extracts the acoustic features for each frame of signals and then compares these values of features with the present thresholds to classify each frame. The pattern-matching method on the other hand needs to estimate the model parameters of speech and noise signal and the detection process is similar to a recognition process. Threshold-based method does not keep much training data and training models and is simpler and faster compared with the pattern-matching method.

The conventional endpoint detection methods are mainly based on the simple energy detector, which performs adequately for clean speech. Most of these methods use short-time energy and zero-crossing as its algorithm for pinpointing the beginning and ending point in a high signal-to-noise condition, but their performance degrades in a noisy environment.

Aerodynamics indicates that the speech signal is non-linear and 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. Compared with the conventional algorithms, our method carry out calculations based only on the time-domain waveform of speech signal, therefore it's low complexity. Experimental results show good performance in extracting the speech segments from utterance with a variety of background noise and a high recognition rate.

II. 본 론

1. 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 in feature, the nonlinear features of speech are often ignored.

(2) Most methods perform well in quite environment, but degrades rapidly in noise ones.

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 locating the beginning and ending point of a speech signal in many previous research work.

Among various endpoint detection approaches, energy-based methods are the mostly widely applied solution to this problem. It has been found that the amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. 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

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

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

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

The short-time zero-crossing rate is too unstable to be used on endpoint detection. It's often accompanied with short-time energy.

Based on the combination of the two features, one such algorithm for locating the beginning and end of a speech signal was proposed by Rabiner and Sambur[6] in the context of an isolated word speech recognition system[7]. Their algorithm is fast and practical since the speech signal is acquired at the same time as the word boundary detection is done, but it couldn't guarantee its success in a noisy environment.

2. Time-dependent Lyapunov Exponents Algorithm

A mass of experiments indicate many physical phenomena present a complex behavior with fluctuations over time. A detailed model of the vocal tract should consider the time variation of vocal tract shape, the vocal tract resonances, losses due to heat conduction and viscous friction at the vocal tract walls, nasal cavity coupling, softness of the vocal tract walls, the effect of subglottal (lungs and trachea) coupling with vocal tract resonant structure and radiation of sound at the lips. A time-varying linear filter can model the effects of some of these factors, but the remaining ones are very difficult to model. Some techniques have been proposed in the literature to analyze the non-linearities of dynamical systems, including those systems where it is not currently possible to represent explicitly by a mathematical model. A set of techniques that can perform this analysis constitutes so-called Chaos theory.

Non-linear dynamical systems theory or just Chaos theory can use time series to characterize the dynamical properties of a system and extract information from these data. Thus, speech production can be analyzed by the techniques underlying this theory, and the information extracted can be applied to improve the accuracy of many speech processing systems.

In this paper, we explore the extraction of an important non-linear dynamical feature, namely time-dependent Lyapunov Exponents

As we know in mathematics, the Lyapunov exponent of a dynamic system is a quantity that characterizes the rate of separation of infinitesimally close trajectories. Consider two points in a space, X_0 and $X_0 + \triangle x_0$, 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

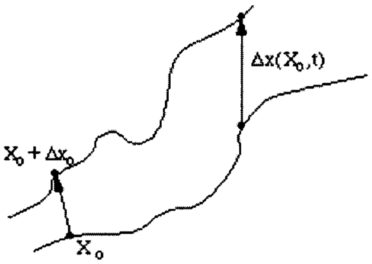


그림 1. 두 점으로부터 궤도 분리
Fig. 1. The separation of two orbits from the two points.

function of time.

According to many researches, if a system is unstable, then the orbits diverge exponentially for a while, but eventually settle down. For chaotic points, the function$\triangle x(X_0, t)$ 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

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

This number, called the Lyapunov exponent "λ", is useful for distinguishing among the various types of orbits. It determines the predictability of a dynamic system. A positive LE is usually taken as an indication that the system is chaotic.

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[8~9] and make some conversion to serve the need of 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 divide 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 into many small sects based on the sample number in the frame. The sketch map was shown in Fig.2.

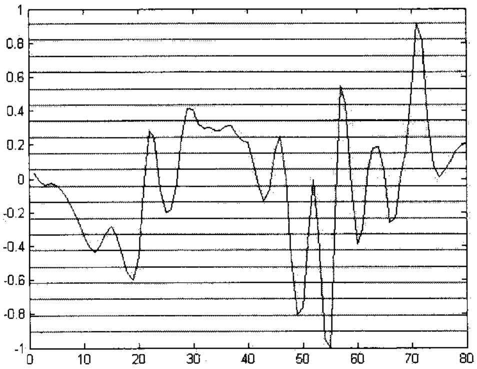


그림 2. 각 프레임의 스케치 맵 분리(가로축은 각 프레임의 샘플링개수)

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 $n \geq 2$, compute the amplitude dispersion d0 d1 d2···dn between two close samples from the first sample in the region;

d. For each of the sample pairs in step c, we look for another couple of points succeeding the current sample pair in the curve. Some of these derived points maybe out of this region or overlap the current sample from step c. Then compute the amplitude dispersion of the new couple of samples d0', d1', d2'···dn';

e. Compute Lyapunov exponent using the following formula:

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

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.

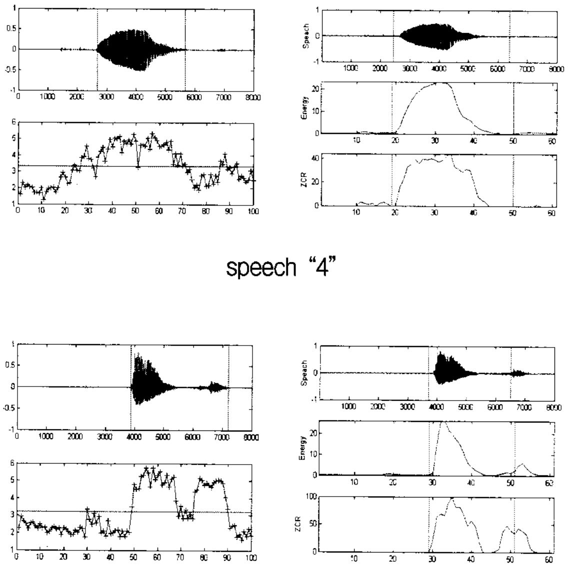
g. Then choose the next frame and repeat the steps mentioned above.

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

III. 실 험

The "GoldWave" software was used to record the set of English numbers 0-9 at a sampling rate of 8 kHz in a laboratory environment. The test subjects consisted of 20 individuals and each of'the subjects recorded two sets of samples. Then we carry out the endpoint detection using the proposed method and compared it with the conventional algorithm.

The comparative results are shown in Figure 3. The left-column outputs were processed using the Lyapunov Exponents method while conventional algorithm was used for the right-column outputs. We can see from the graphs in the right-column that the conventional detection method will miss some parts of the speech or include noise as a part of the speech if the threshold of energy or zero-crossing rate is not properly set. Contrast this with our proposed method where we have a better result since we could easily select an optimal threshold based on the more accurate separation of the speech segment from the noise.



speech "4"

speech "8"

그림 3. 음성으로부터의 종점 탐색 결과(왼쪽 부분은 리아프노프 지수 방법, 오른쪽- 기존 방법)

Fig. 3. The endpoint detection results on "clean speech". Left-column used the Lyapunov exponents method while the right-column used conventional algorithm.

표 1. 잡음 환경하의 제안한 방법의 출력 결과

Table 1. Output using proposed method under noisy environments.

(Unit: Sampling number)

	Clean Environmen t		Noisy Environment							
			white noise		f-16 cockpit		pink noise		babble noise	
	start	end	start	end	start	end	start	end	start	end
0	2200	6240	2480	5760	2440	5520	2280	5520	2440	6240
1	1400	4240	1480	4160	1400	4160	1400	4080	1400	4160
2	2440	5840	2440	5760	2440	5840	2440	5840	2440	5760
3	2600	6320	2840	6400	2680	6320	2600	6240	2600	6240
4	2680	5360	2840	5360	2680	5520	2680	5360	2680	5520
5	2840	5600	2680	5520	2840	5520	2840	5440	2840	5520
6	3080	7440	3240	7440	3080	7440	3000	7440	3000	7440
7	3000	6800	3000	6960	2920	6560	3000	6880	2920	6880
8	3960	7200	3960	7040	3960	7120	3960	7120	3960	7200
9	2680	6080	2840	6000	2760	5920	2760	6000	2680	6080

표 2. 인식 정확도

Table 2. Recognition accuracy.

	Recognition Accuracy (%)	
Speech	new endpoint detection	conventional method
1	100.0	100.0
2	97.5	95.0
3	100.0	100.0
4	97.5	90.0
5	95.0	92.5
6	100.0	95.0
7	97.5	92.5
8	100.0	100.0
9	95.0	90.0
0	100.0	100.0

The noise samples from noise database NOISEX92 were utilized as additional noise to each sample set in order to evaluate the performance of the proposed method simulating speech production under various noisy conditions. We chose the representative noise, white noise, F-16 cockpit noise, pink noise and babble noise. The results are shown in Table 1.

We can see from Table 1 that although some endpoints are affected by the introduced noise, the error is less than 256 sampling points or one frame. This small error confirms that the proposed endpoint detection method using Lyapunov Exponents performs effectively even under noisy conditions.

Less but more typical parametric characteristics for pattern matching are then extracted based on the accurate endpoint detection method. As the recognition results in Table 2 shows, the average recognition accuracy has been improved to 98.25%.

Ⅳ. 결 론

Digital isolated word recognition system requires very high accuracy in locating the beginning and end points of a speech in which endpoint detection plays a crucial part. Conventional algorithms based on the short-time energy and zero-crossing rate performs adequately for clean speech but fails in adverse or noisy environments. Several research studies have been done lately on the characterization of the speech signal using non-linear dynamical features and our proposed algorithm of Lyapunov exponents as a method of endpoint detection exploits this approach as a way of improving the accuracy of speech recognition.

Simulation results showed that the proposed method could extract or separate the speech segment from the noise segment more precisely and with less complex computational task than conventional algorithms or methods. Even the introduction of various noise into the speech sample set did not have a major degradation on its performance but actually proved the effectivity of our proposed method even under various noisy conditions.

Our proposed endpoint detection method using Lyapunov Exponent which was able to increase the average recognition rate up to 98.25% for isolated words on unspecified person can be a valuable tool for speech recognition systems in the future.

참 고 문 헌

[1] Woo-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee, Jong-Seok Lee, "Speech/non-speech Classification using Multiple Features for Robust Endpoint Detection", International Conference on Acoustics, Speech, and Signal Processing, 2000.

[2] Stefaan Van Gerven, Fei Xie, "A Comparative Study of Speech Detection Methods", European Conference on Speech, Communication and Techonlogy,1997.

[3] Ramalingam Hariharan, Jula Hakkinen, Kari Laurila, "Robust End-of-utterance Detection for Real-time Speech Recognition Applications", International Conference on Acoustics, Speech, and Signal Processing, 2001.

[4] A. Acero, C. Crespo, C. De la Torre, J. Torrecilla, "Robust HMM-based Endpoint Detector", International Conference on Acoustics, Speech, and Signal Processing, 1994.

[5] E. Kosmides, E. Dermatas, G. Kokkinakis, "Stochastic Endpoint Detection in Noisy Speech", SPECOM Worshop, 109-114, 1997.

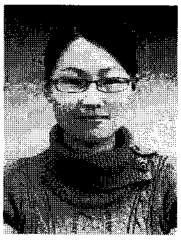
[6] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., vol. 54, No. 2, pp. 297-315, February 1975.

[7] 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.

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

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

저자소개



장 한(학생회원)
2007년 중국 회혜공과대학 공정 장비제어학과 학사졸업.
2009년 현재 전북대학교 전자정보 공학과 석사과정.
<주관심분야 : Mechanical Engineering, Automotive Design, Electronic Control Technology, Signal Processing>

정 길 도(정회원)-교신저자
1984년 Oregon State University 기계공학 학사졸업.
1986년 Georgia Institute of Technology 기계공학 석사졸업.
1992년 Texas A&M University 기계공학 박사 졸업.
2008년 현재 전북대학교 전자정보공학과 교수
<주관심분야 : Time-Delay, Robotics, 인공지능, Web 기술>