ENDPOINT DETECTION IN NOISY ENVIRONMENT USING A POINCARÉ RECURRENCE METRIC

Lingyun Gu, Jianbo Gao and John G. Harris

Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL, 32611, U.S.A.

Abstract

Speech endpoint detection continues to be a challenging problem particularly for speech recognition in noisy environments. In this paper, we address this problem from the point of view of fractals and chaos. By studying recurrence time statistics for chaotic systems, we find the nonstationarity and transience in a time series are due to non-recurrence and lack of fractal structure in the signal. A Poincaré recurrence metric is designed to determine the stationarity change for endpoint detection. We consider the small area of beginning and ending of an utterance as transient. For nonstationary and transient time series, we expect the average number of Poincaré recurrence points for each given small block will be different for different blocks of data subsets. However, the average number of recurrence points will stay nearly constant. The resulting recurrence point variability algorithm is shown to be well suited for the detection of state transitions in a time series and is very robust for different types of noise, especially for low SNR.

1. INTRODUCTION

The detection of the endpoints of an utterance is required in many speech applications. Accurate endpoint detection is crucial for good speech recognition accuracy. The most popular existing detection method is the simple energy detector which performs adequately for clean speech. Problems arise in noisy environments for low energy phonemes (some fricatives and plosives, for example) at the endpoints. A major source of error in isolated word speech recognition systems is the inaccurate detection of the beginning and ending boundaries of test and training patterns. The performance of existing endpoint detection severely degrades in noisy environments. Generally speaking, the incorrect determination of endpoints for an utterance results in at least two negative effects [8]:

1. Recognition errors are introduced;
2. Computation increase.

The types of errors for energy-based detectors introduced by poor SNR include [7]:

1. Missing the leading or trailing low-energy sounds such as fricatives;
2. Classifying clicks, pops and other background noise as part of speech due to their high energy content;
3. Falsely classifying background noise as speech while missing the actual speech. This is particularly true when the background noise consists of speech from other speakers, such as in babble noise.

All of the above errors in turn reflect negatively on the overall performance of the recognition system. To deal with these problems, many advanced algorithms have been proposed during the past decades [3, 4, 5, 6, 7, 8, 9]. Some of these algorithms combine several existing good-merit features [7, 8, 9] while others utilize new features [3, 4, 5, 6]. Although better results are achieved, the performance of all these algorithms severely degrades as the SNR decreases.

In this paper, we propose a new algorithm based on the theory of fractals and chaos, which is widely used in linear and nonlinear time series analysis techniques. Here, the average number of Poincaré recurrence points (defined in the next section) for each designed sliding block for the waveform is considered as a new feature [1, 2]. After the characteristic curve is drawn, an adaptive threshold is set to determine the correct endpoints based on simple signal modeling.

The proposed algorithm provides several advantages:

1. High accuracy (see results in Section 4);
2. Performance does not degrade too severely with increasing levels of noise.
3. There is no need to estimate the background noise as is commonly required for other endpoint detection algorithms.

The paper is organized as follows. The new algorithm is derived in Section 2 and the whole system design is proposed in Section 3. In Section 4, algorithm performance is quantified under different levels and types of noise backgrounds. Finally, our conclusions are summarized in Section 5.

2. NEW ALGORITHM DESCRIPTION

In dynamics, most methods for detection of non-stationarity are based on quantifying features of nearest neighbors. The nearest neighbors are also called Poincaré recurrence points, and are further divided into two classes, with two types of recurrence times. Given a scalar time series $\{x(i), i = 1, 2, \ldots\}$, we first construct vectors of the form: $X_i = [\{x(i), x(i+L), \ldots, x(i+(m-1)L)\}]$, with m being the embedding dimension and L the delay time. $\{X_i, i = 1, 2, .., N\}$ then represents a certain trajectory in m-dimensional space. In this paper, we shall always normalize the time series into the unit interval $[0, 1]$ before subsequent analysis. Next, we arbitrarily choose a reference point X_0 on the reconstructed trajectory, and consider recurrences to its neighborhood of radius r: $B_r(X_0) = \{X : \|X - X_0\| \leq r\}$. The subset of the trajectory that belongs to $B_r(X_0)$ is denoted by $S_1 = \{X_{t1,} X_{t2,} \ldots, X_{ti,} \ldots\}$. The elements of the set S_1 are the Poincaré recurrence points. Using S_1, we define the Poincaré recurrence time as the element of $\{T_1(i) = t_{i+1}-t_i, i = 1, 2, \ldots\}$. For later convenience, we call the elements of $\{T_1(i)\}$ the recurrence times of the first type. Sometime we may have $T_1(i) = 1$ (for continuous-time systems, this means 1 unit of sampling time), for some i. This corresponds to both X_{ti} and X_{ti+1} belonging to S_1. For deterministic continuous-time systems with fixed sampling time, if the radius r of $B_r(X_0)$ is not too small, then we can have a sequence such as X_{ti}, X_{ti+1}, …, X_{ti+k} belonging to S_1, with $k >> 1$. Figure (1) shows this schematically. We call the points $X_{ti+1,}$ …, X_{ti+k} (excluding X_{ti}) "sojourn points". When $k >> 1$, each such sequence of points effectively represents a one-dimensional (1D) set. For maps or continuous systems with small r, the number of sojourn points are negligible. Hence, sojourn points form a 0D (empty or almost empty) set. We now remove these points from S_1 by $S_2 = \{X_{t_1^{'}}, X_{t_2^{'}}, \ldots, X_{t_i^{'}}\}$, which in turn defines a time sequence $\{T_2(i) = t_{i+1}^{'} - t_i^{'}, i = 1,2,\ldots\}$. We call the elements of S_2 recurrence points of the second type, and $T_2(i)$ recurrence times of the second type.

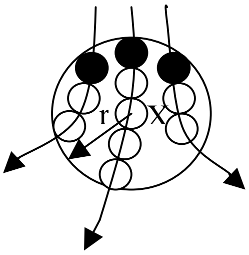

Figure 1: A schematic showing the recurrence points of the second type (solid circles) and the sojourn points (open circles) in $B_r(X_0)$ [1, 2]

For dissipative chaotic systems, we have shown that with fixed r, the distribution of $\{T_2(i)\}$ is exponential, due to the memoryless property of chaotic systems, and the mean of $T_1(i)$ and $T_2(i)$ are both related to the information dimension d_1 of the attractor by simple scaling laws:

$$\overline{T}_1(r) \propto r^{-d_1}$$

$$\overline{T}_2(r) \propto r^{-d_1^{'}}$$

with $d_1' = d_1$ for discrete maps and continuous-time systems with small r (when the sojourn points form a 0D set), and $d_1' = d_1-1$ for continuous-time systems with large r (when the sojourn points form a 1D set). For a periodic signal, $T_2(i)$, this simply provides an estimate of the periodicity of the signal. Based on an observation that, due to non-stationarity, successive recurrence times of the second type will, on average, be changing with time. So we design the following algorithm to detect nonstationarity and state transitions.

For the given utterance, partition a long-time series into overlapping blocks of data sets of short length k, and compute $\overline{T}_2(r)$ for each data subset. The length of the subset is chosen to be short enough so that non-stationarity is not a problem for the subset. At the same time, the subset is long enough so that $\overline{T}_2(r)$ can be reliably estimated. Usually, overlapping blocks are preferred so that bifurcation can be more accurately located. For nonstationary and transient endpoint area, we show that $T_2(r)$ is different for different blocks of data subsets.

3. SYSTEM DESIGN

3.1 Pre-emphasizing the Input Signal

The input signal is first filtered with a bandpass filter from 250Hz to 3750Hz (FIR filter of order 50). This band, very similar to the band of telephone lines, is generally considered to contain the most overall speech information. Thus this type of fixed filtering is reasonably effective for improving the signal to noise ratio of speech to non-speech.

3.2 Calculate the Characteristic Curve of $\overline{T}_2(r)$

From the beginning of an utterance, it is partitioned into overlapping blocks. Here, we set each block to contain 1000 sample points with a 100-sample overlap. Based on the parameter introduction in Section 2, the embedding dimension *m* is set to 4 and the delay time *L* is set to 50 sample points. For each available computational sample point, we obtain a single $T_2(r)$. Then $\overline{T}_2(r)$ for one block is calculated by taking the average of all the single $T_2(r)$ in that block. It is easier to understand this step by comparing the standard method of block-based energy curve.

3.3 Crudely determine the transient area

After we get the characteristic curve, we first set two hard thresholds *Thresh1* and *Thresh2* for the first beginning and final ending, respectively. Subsequently, two corresponding windows (Window1 and Window2) are defined. Here, we consider the portion of the utterance after Window1 and before Window2 must be voiced. Then we average the value of all the windows from the first to Window1 to get the base line for the beginning area and similarly for the ending area.

3.4 Final Endpoint Detection

In the beginning area, we search forward to the end from the first window, if there exists a peak or valley, whose difference with the base line is greater than a given *Thresh3*, we consider it is the beginning window. Similarly, we search backwards to the beginning from the last window. Then we will get the ending window by comparing with *Thresh4*. Looking back to the original waveform, we define the center of the located windows as the true endpoints. Figure (2) shows an explicit example for the word "zero".

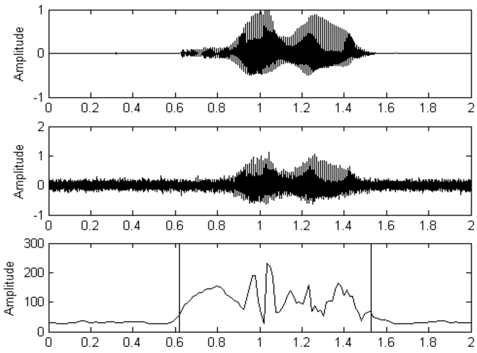

Figure 2: Top, the clean speech; Middle, white noise case at 5dB; Bottom, the recurrence characteristic curve

4. EXPERIMENTS AND RESULTS

4.1 Database and Noise Addition

The experiments have been run on a total of 600 isolated English digits (6 men and 6 women) from the TI46 database. Each speaker pronounced digits zero to nine five times. All of the utterances were manually labeled before the experiments. The sampling rate of the database is 12500 Hz. For generating the noisy speech files, we used different types of noise available from the Signal Processing Information Base (SPIB) collected by Rice University [10]. Three sorts of noise are considered in this paper: white noise, pink noise and babble noise. The sampling rate of the noise database is 19980 Hz but was downsampled to 12500 Hz to match the TI46 database. To set up the noisy speech database for testing, we added the prepared noisy signals to the recorded speech signals with different SNRs including 5, 10, 15 and 20 dB. The speech waveform and the noise waveform are each calibrated and the noise segment is randomly selected from the noise file.

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. Here, we choose the first option for the most straightforward comparison. We consider that an endpoint is lost if the error is higher than 75 ms for the beginning and 100 ms for the end.

4.2 Experimental results

Figures (3), (4) and (5) show the system performance by providing different noise source of white, pink and babble and comparing the algorithm performance with that of the pure energy endpoint detector. The plots show the accuracy in three components for each algorithm: beginning, ending and overall performance. We see that the new algorithm's performance is significantly better than the pure energy detector. The new algorithm is also not sensitive to the degradation of noise, even in the lower SNR, though the performance degraded much below 10 dB. For the white noise case, the accuracy increases from 53.1% to 85.3% at 5 dB, while increasing from 39.3% to 82.3% and from 36.8% to 77.3% for the pink noise and babble noise at 5 dB, respectively.

5. CONCLUSIONS

The new algorithm for isolated words endpoint detection has been proposed in this paper. This algorithm introduces a new idea from fractal and chaos, to achieve excellent overall results. In particular this method is able to reliably detect the onset and offset of speech even for weak beginnings and endings. The major benefit of the algorithm is that the endpoint can be detected in the presence of different kinds of noise that have much greater energy than the initial and final speech segments.

6. REFERENCE

[1] J.B. Gao, *Detecting Nonstationarity and State Transitions in a Time Series*. In Physical Review E. Vol. 63, pp.0662021 – pp. 0662028, May, 2001

[2] J.B. Gao, *Recurrence Time Statistics for Chaotic Systems and Their Applications*. In Physical Review Letters, Vol. 83, No.16, pp. 3178 – pp. 3181, Oct, 1999

[3] Jean-Claude Junqua, Brian Mak, and Ben Reaves, *A Robust Algorithm for Word Boundary Detection in the Presence of Noise*. In IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 3, pp. 406 – pp. 412, Jul, 1994

[4] Qi Li, Jingsong Zheng, Augustine Tsai and Qiru Zhou, *Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition*, In IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 3, pp. 146 – pp.157, Mar, 2002

[5] Chin-Teng Lin, Jiann-Yow Lin and Gin-Der Wu, *A Robust Word Boundary Detection Algorithm for Variable Noise-Level Environment in Cars*. In IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, pp.89 – pp.100, Mar, 2002

[6] J. Navarro-Mesa, Moreno-Bilbao and E. Lleida-Solano, *An Improved Speech Endpoint Detection System in Noisy Environments by Means of Third-Order Spectra*. In IEEE Signal Processing Letters, Vol. 6, No. 9, pp. 224 – pp. 226, Sep, 1999

[7] 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, pp. 3808 – pp. 3811, 2002

[8] Lingyun Gu and Stephen A. Zahorian, *A New Robust Algorithm for Isolated Endpoint Detection*. In Proc. IEEE ICASSP - 02, Vol. 4, pp. 4161 – pp. 4164, 2002

[9] Liang-Sheng Huang, Chung-Ho Yang, *A Novel Approach to Robust Speech Endpoint Detection in Car Environments*. In Proc. IEEE ICASSP-00, pp.1751-pp.1754, 2000

[10] http://spib.rice.edu/spib/select_noise.html

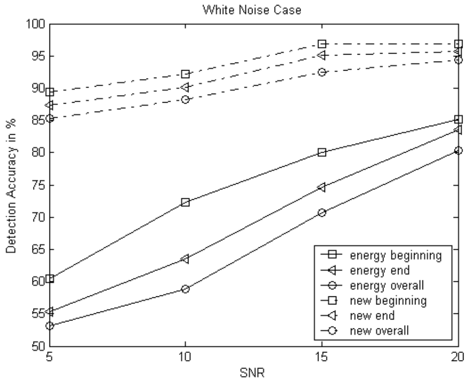

Figure 3: White noise case in terms of SNR: 5-20dB

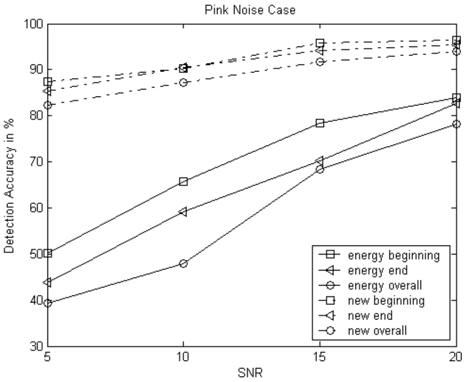

Figure 4: Pink noise case in term of SNR: 5-20dB

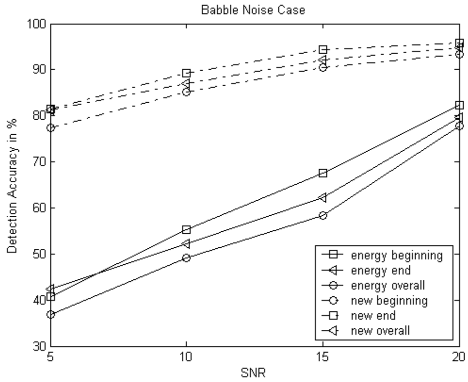

Figure 5: Babble noise case in term of SNR: 5-20dB