Fractal Characteristic-Based Endpoint Detection for Whispered Speech*

XUEQIN CHEN, HEMING ZHAO
School of Electronics and Information Engineering
Soochow University
No.178 GanJiang Dong Lu, SuZhou, 215021
P.R.CHINA

Abstract: - In this paper, a fractal based approach is proposed to detect endpoints in whispered speech. The underlying principle is based on the fact that whispered speech is sufficiently chaotic and thus can be analyzed using fractal theory. Due to the different scope of fractal dimensions of silence, noise and speech segment, speech/non-speech segment could be determined from that with a simple decision scheme. The results of experiments, which have been performed on two databases and different methods, show the efficiency of the proposed whispered speech segmentation algorithm.

Key-Words: Endpoint detection; Fractal; Fractal Dimension; Robust; Whispered speech

1 Introduction

Endpoint detection of speech signal is a necessary step of speech process in speech recognition, speech coding, speech enhancement, mobile and multimedia communication. Presently at the quiet circumstance just like laboratory, endpoint detection of speech has achieved a high accuracy. But at the low SNR circumstance, a lot of studies still focus on this issue[1-3].

Whisper is a common form of speech communication way that one uses for a variety of reasons. For example, with the mobile telephone used widely, people often need to call in a whisper at public for avoiding disturbing others and assuring the secrecy of communication. For the special pronouncing mechanism, whispered speech has no vocal fold vibration and its energy is much lower than that of normal speech. So suitable feature need to be designed to this question. It is pity that research reports about this field are very few presently. To our best knowledge, only reference [4] gives an endpoint detection algorithm based on entropy for whispered speech and its mean correct rate is 97.6%.

As view from the waveform characteristic of whispered speech, fractal characteristic coming of graphics is used to detect endpoints of whispered speech signals in this paper and the detailed algorithm is also listed. Experiments show that the correct rate of this method could achieve 99%.

* Supported by the National Natural Science Foundation of China (Grant No. 60572076) and the University Natural Science Research Project of Jiangsu Province (Grant No. 05KJB510113)

2 Fractal and Endpoint Detection of Whispered Speech

For a given input signal, the fractal dimension of each frame is estimated firstly. Boundaries of speech segments are then detected based on the fractal dimension by a decision rule. The block diagram of the proposed method is depicted in Fig.1.

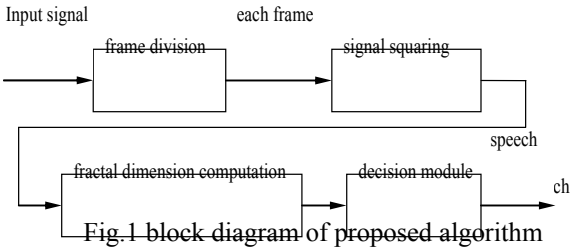

Fig.1 block diagram of proposed algorithm

2.1 Theory Basis of Fractal Applied to Whispered Speech

A Chinese syllable is composed of an initial segment which is usually consonant and a final segment which is often acted by simple or compound vowel, sometimes with a terminal “n” or “ng”. As viewed from pronouncing mechanism, consonant is formed by turbulent flow created by the exhaled air stream obstruction by glottal and lip with no vocal cords vibration, and differently vowel is the sound created by the relatively free passage of breath through the larynx and oral cavity where the air from the lungs cause the vocal cords of the larynx to vibrate[5].

In initial segment, the pronunciation of whispered speech and normal speech is similar. But that is

different in the final segment. In whispered speech, the glottis is half-open which means the forepart of glottis is close and the back-end of it is a triangle cranny. The lower pair of vocal cords doesn't vibrate, air stream from lung produce friction noise through the open part of glottis. Exhaled air stream forms turbulent flow by glottal constriction. So the initial and final part of whispered speech all could be considered a turbulent flow. Turbulent flow has been proved a typical chaos. Furthermore, acoustic and air dynamics have proved that speech signal is a complex nonlinear process which have chaos mechanism. Here the fractal theory adopted by this paper is just an effective mathematic tools describing for chaos signal[6,7].

2.2 Fractal Dimension

A fractal is an object or quantity that displays self-similarity of nature, in a somewhat technical sense, on all scales. The object need not exhibit exactly the same structure at all scales, but the same "type" of structures must appear on all scales. There are multi-parameter to describe fractal character where fractal dimension is a parameter in common use. Fractal dimension is an important characteristic of signal which can keep the information of how complex the signal construction is [6]. So fractal dimension is often used to analyze noised speech signal.

To different question, the definition of fractal dimension is different. In this paper, box-counting dimension D_B is chosen to be fractal characteristic for its easy-computing and efficiency. Supposing that there is a graphic G, covered with a grid which side length of square is ε. There will be $N_G(\varepsilon)$ squares intersecting with the graphic. The box-counting dimension of G could be defined as

$$D_B = \lim_{\varepsilon \to 0} \frac{\ln N_G(\varepsilon)}{\ln(\varepsilon)} \tag{1}$$

The number of box-counting scale M is a key parameter to result. In this paper, M is selected based on a large number of experiments and an experiential formula[8]:

$$M_x = M_y = 1.87 \times (N-1)^{\frac{2}{5}} \tag{2}$$

Where M_x, M_y represent the number of box-counting scale of X and Y axis respectively. N is the number of sampling points of this frame.

2.3 Algorithm of Extracting Fractal Characteristic

The fractal dimension adopted here is the box-counting dimension described in 2.2. For the purpose of avoiding the influence of signal amplitude, the signal will be squared firstly. That is to compress or extend the amplitude scope of the signal to make each frame signal be a square. So the fractal dimension of signal will be independent of signal amplitude and be related to waveform of signal only. That is just one important advantage of fractal theory applied for speech endpoint detection. Then fractal dimension value could be computed based on squared signal. The detailed computing steps of a certain frame are listed as follows.

step1. Choosing M, that is the number of box-counting scale. It means that each frame signal will be computed M times with different box-counting scale. ($M_x = M_y = M$).

step2. For each scale, defining the side length of square, $L(i)$, $i = 1 \sim M$.

step3. Using the gridding (which is composed by $\frac{N}{L(i)} \times \frac{N}{L(i)}$ squares) to cover the squared signal. There N is the number of sampling points of this frame.

step4. Computing the number of the needed least squares intersecting on this frame signal. That is the box-counting number $N(i)$ of this scale $L(i)$ for this frame.

step5. If $i < M$, $i = i + 1$, repeat steps 1~4 until $i = M$, thus a set of $N(i)$ could be got, $i = 1 \sim M$.

step6. Computing the each value of

$$A(i) = \log\frac{1}{L(i)}, \quad B(i) = \log\frac{N(i)}{L(i)}, \quad i=1\sim M \tag{3}$$

step7. Then using the least square method (LSM) simulating the line $A(i) \sim B(i)$, then the slope of this line is the box-counting dimension D of this frame speech signal.

$$D = \frac{(\sum_{i=1}^{M} B(i))(\sum_{i=1}^{M} A(i)) - M\sum_{i=1}^{M} A(i)B(i))}{(\sum_{i=1}^{M} A(i))^2 - M(\sum_{i=1}^{M} A(i)^2)} \tag{4}$$

We know that M is an important value to fractal dimension. At the basic of formula(2) and many experiments this paper choose M equal 20 which is apropos for whispered speech signal. After computing the dimension D, 9-point median filtering

is used to smooth the fractal dimension curve for the followed endpoint detection.

2.4 Decision Schedule

According to the aforementioned algorithm the fractal dimension D of each frame has could be computed. Usually the fractal dimension of silence is 1.5, and that of white noise is about 2.0 and that of speech ranges from 1.6 to 1.8. Based on the pronouncing mechanism, whispered speech signal is completely noise exited and there is no vocal cords vibration. The result in speech waveform is that characteristics are similar between initial and final segment at time domain. So the fractal dimension of a length of speech signal will maintain a steady level. This characteristic is helpful for choosing threshold value to judge the endpoint and it is also an important reason why the fractal algorithm is suitable for endpoint detection of whispered speech.

The fractal dimension is apparent different between speech and non-speech segment. So the decision rule is not complex. Firstly, we can find the maximum $D_{\max}$ and minimum $D_{\min}$ of fractal dimension and compute the mean value D_{ave} of them.

$$D_{ave} = \frac{D_{\max} + D_{\min}}{2} \tag{5}$$

Then D_{ave} will be the threshold of the first time decision. Commonly, the exact location is near by the primary location. But second time decision based some rules is necessary for sporadic false dots are inevitable. Based on the primary location we can search forwards and backwards to find the exact location and pick the speech segment.

3 Experiment Results

In order to evaluate the performance of the proposed approach, we carried out a number of evaluation tests. The whispered speech data are general data from the Whisper_N database (the whispered speech database constructed by the researchers of Nanjing University [9]) and Whisper_S database (the whispered speech database constructed by the researchers of Suzhou University) with a 2~10dB signal-to-noise relation (SNR) from different male and female speakers, 380 Chinese syllable composed by 21 initial consonants and 35 final segments(covering most Chinese syllables). Signal, sampled at 11025Hz, is divided into frame, 23.2 ms long.

Forwards accordingly, M is a key parameter to fractal dimension. Fig.2 shows that the different fractal dimension curve based on different M. If M is too small, the result warps badly. When M is bigger than 20 the result almost doesn't change with M. So the effective M for whispered speech endpoint detection is 20.

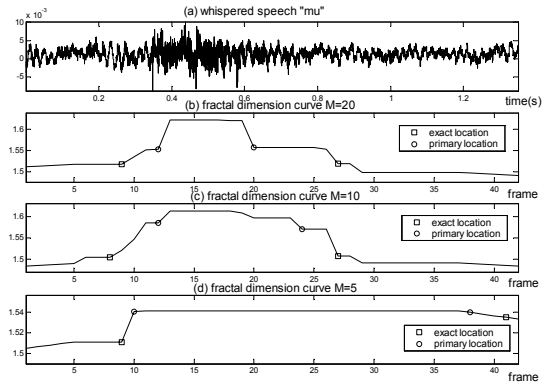

Fig.2 different fractal dimension curve corresponding M

Then 9-point median filter is used to smooth firsthand box-counting dimension preparing for endpoint detection. Fig.3 shows the fractal dimension curve before filtered and after filtered. We can see from Fig.3 that most of the curve burrs have been smoothed after filtered, which is very helpful for endpoint location.

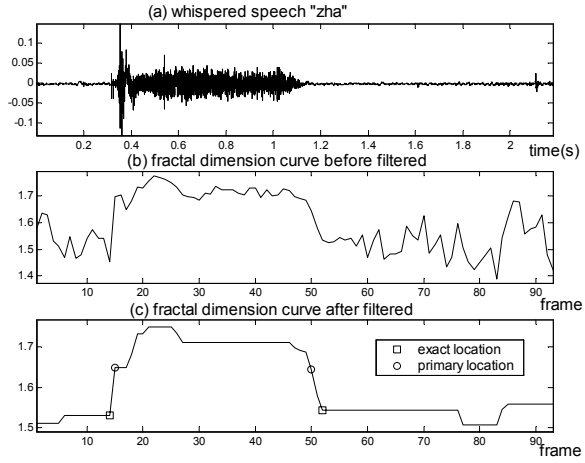

Fig.3 the fractal dimension curve before filtered and after filtered

From the observation in Fig.2, we can find that the SNR is very low. It is difficult even for human eye to divide its endpoint but the method proposed by this paper could pick up the speech segment from signal stream effectively. If strengthening the noise (for example, adding a white noise to the whispered speech) the result is still satisfying. Fig.4 is an example of endpoint detection of noised whispered speech where the noise is white noise. The fractal dimension of white noise is bigger than that of whispered speech, so despite of the opposite curve

between (b) and (d) in Fig.4, the location is still distinct.

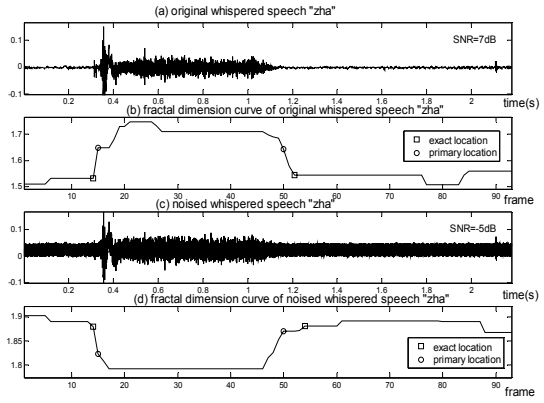

Fig.4 endpoint detection of noised whispered speech (where the noise added is white noise)

The performance of the proposed algorithm is also compared with the method proposed by reference [4] on two different databases (each 1900 samples of whisper_N and whisper_S, covering most Chinese syllables). The correct rates are summarized in Table1.

Table1. The correct rates of the algorithm proposed by this paper and reference [4] on different whispered databases

database \ method		Entropy	Fractal
Whisper_N (1900samples)	Correct number	1854	1882
	Correct rate (%)	97.6	99.05
Whisper_S (1900 samples)	Correct number	1863	1886
	Correct rate (%)	98.1	99.26
Mean correct rate (%)		97.85	99.15

The errors presented in experiments mainly concentrate on some syllables beginning with “m” and some syllables ending with “eng”. The scope of error accepted is one frame.

4 Conclusion and Discussion

Whispered speech has the characteristics of low energy and low SNR for its special pronunciation, so the conventional endpoint detection algorithms designed for normal speech become non-effective for whispered speech. With a view to the chaos phenomenon of whispered speech, this paper achieves whispered speech endpoint detection based on fractal dimension theory and gives the detail algorithm steps. It is recommendable that this method is independent of speakers and contents. Experiments show that this method is very effective and robust. So this method could be used to the pre-process of recognition and reconstruction of whispered speech[10,11]. Further, it is hoped that the feature could be improved to be used to detect initial and final segmentation.

References:

[1] M.Pwint, F.Sattar, A New Speech/non-speech Classification Method Using Minimal Walsh Basis Functions. *IEEE International Symposium on Circuits and System*. Vol.3, 23-26, 2005, pp.2863-2866

[2] B.Y.Smolenski, U.O.Ofoegbu, K.S.Jashmin, E.Y.Robert, Nonlinear State Space Embedding Features and their Application to Robust speech Segmentation. *Proc. ISPACS*, 2004, pp.317-320

[3] B.F.Wu, K.C.Wang, Robust Endpoint Detection Based on the Adaptive Band-partitioning Spectral Entropy in Adverse Environments. *IEEE Transactions on Speech and Audio Processing*. Vol.13, No.5, 2005, pp.762- 775

[4] X.L.Li, H.Ding, B.L.Xu, Entropy-based Initial/Final Segmentation for Chinese Whispered Speech. *Acta Acustica*, Vol.30, No.1, 2005, pp. 69-75

[5] Z.J.Wu, M.S.Lin, *Introduction of Experimental Phonetics to Chinese*, Higher Education Press, 1989

[6] M.A.Akaidi. *Fractal Speech Processing*, Cambridge University Press, 2004

[7] D.Saupe, H.Jurgens, H.O.Peitgen, *Chaos and Fractals*, Springer Verlag, 1992

[8] R.Sengupta, N.Dey, D.Nag. Comparative Study of Fractal Behavior in Quasi-random and Quasi-periodic Speech Wave Map. *Fractals*, Vol.9, No.4, 2001, pp.403-414

[9] L.L.Yang, Y.Li, B.L.Xu, The Establishment of A Chinese Whisper Database and Perceptual Experiment. *Journal of Nanjing University* (*Natural Sciences*), Vol.41, No.3, 2005, pp.311-317

[10] T.Toh, K.Takeda, F.Itakura, Analysis and Recognition of Whispered Speech. *Speech Communication*, Vol.45, No.2, 2005, pp.139-152

[11] R.W.Morris, M.A.Clements, Reconstruction of Speech From Whispers. *Medical Engineering & Physics*, Vol.24, 7-8, 2002, pp.515-520