Speaker Verification Based Fractal Geometry

Loay E. George, *Laith A. Al-Ani, *Mohammed Sahib Mahdi

Department of Astronomy, College of Science, University of Baghdad, Baghdad-Iraq.
** Department of physics, College of Science, University of Al-Nahrain. Baghdad-Iraq.*

Received:27/10/2002 Accepted:31/10/2005

Abstract

The research shows that the fractal structure of the speech signal exhibits fractal characteristics. The encouraging analysis results indicated that the fractal dimension have good discrimination capabilities for speech signal, where it gave a speaker recognition percentage about 70% (In the field of research for a single parameter this percentage be valued), and these capabilities are strongly depends on the setting situations of the recording process. The pitch period is the another used parameter supported to the fractal dimension to strengthen the verification decision. Where the pitch period estimated by a new suggested simple method to gave a speaker verification percentage about 90%. The coalition work of the two parameters (fractal dimension and pitch period) gave a speaker verification percentage of about 85%.

الخلاصة

اظهر البحث إن البنية الكسورية لاشارة الكلام تأخذ تماما شكل كسوري. نتائج التحليل كانت مشجعة، حيث أعطت انطباعا جيدا على قدرات التمييز التي يوفرها البعد الكسوري Fractal dimension لاشارة الكلام، حيث أعطى البعد الكسوري نسبة تمييز متكلمين بحدود 70% (ضمن مجال البحث فان هذه النسبة تعتبر قديرة لأنها تخص عامل تمييز واحد). أيضا وجد إن قدرات تمييز البعد الكسوري تعتمد بشدة على حالات ضبط عملية تسجيل الكلام.

استخدم عامل تمييز ثاني هو نغمة الصوت Pitch periodوالذي اقترن مع البعد الكسوري لتقوية قرار التمييز ضمن حالات اشمل. نغمة الصوت حسبت باستخدام طريقة بسيطة جديدة مقترحة لتعطي نسبة تمييز متكلمين بحدود 90%.

العمل الائتلافي لكل من عاملي التمييز (البعد الكسوري ونغمة الصوت) مهم في تمييز الأصوات المختلفة بنسبة تمييز متكلمين بحدود 85%.

Introduction

Mandelbrot's fractal geometry provides a new qualitative and mathematical approach for understanding the complex shapes of nature. Many objects and patterns in the nature world possess the quality of self-similarity, the magnification portion of the shape look qualitative like the original pattern [1].

The fact that such complicated, and seeming random, shapes of nature can be characterized by a single number, i.e. the fractal dimension D, such that it can be motivate to test, or fractal characterization to different natural signals like the speech wave. The speech waveform structure is highly irregularity shaped signal, which can be treated as a fractal and studied using fractal mathematics [2].

A fundamental concept of the fractal geometry is the fractal dimension, which unlike the topological dimension may also assume non-integer values. Another important property, often uncounted in fractals, is the self-similarity or, self-affinity.

A fractal set which is invariant under a transformation in which all the coordinates are scaled down by ratio r_1, r_2…rn not all equal is said to be self-affine.

One of the most popular methods to estimate the fractal dimension is the box counting method, despite of its complexity it is the accurate. According to this method [3]:

Consider a bounded set $\boldsymbol{A}$ in Euclidean n-space the set $\boldsymbol{A}$ is said to be self-similar, when $\boldsymbol{A}$ is the union of N distance copies of itself, each of which has scaled down by a ratio r in all coordinates. The fractal or similarity dimension of $\boldsymbol{A}$ is given by the relation,

$$Nr^{D} = 1 \quad \text{or,} \quad D = \frac{\log N}{\log \frac{1}{r}} \qquad (1)$$

Suppose one can cover the set $\boldsymbol{A}$ with n-dimensional boxes of size L_{max}. If the set is scaled down by a ratio r, then there are $N = r^{-D}$ subsets, and so the number of boxes of size $L= r.\ L_{max}$ needed to cover the whole set is given by…

$$N(L) = \frac{1}{r^{D}} = \left[\frac{L_{\max}}{L}\right]^{D} \qquad (2)$$

The simplest way to estimate D from (2) is to divided the n-dimensional space into a grid of boxes with side length L and to count the number of non-empty boxes. If $N(L)$ is computed for several values of L, then D can be estimated as the slope of a least squares linear fit of the data $\{\ln(L), -\ln(\mathrm{N(L)})\}$ [4].

Human Speech

The dominant frequency components which characterizes the phoneme, correspond to the resonant frequency components of the vocal tract are named "formant frequency" each syllable has three to five typical iterated formant frequencies that distinguish it from others for the same speaker [5].

The pitch period refers to the fundamental formant frequencies of such vibration or the inherent periodicity in the speech signal. The pitch detectors are computer algorithms applied directly on the speech signal. Mostly it yields a voicing decision as part of their processing, in which up to four classes of speech is distinguished: voiced, unvoiced, combined (e.g. /z/), and silence (e.g. /h/). Since the majority of excitation of the vocal tract for each pitch period occurs when the vocal cords are closed to each other, each period tends to start with high amplitude (referred to as an epoch), and then follow decaying-amplitude envelope. The rate of the decay is usually inversely proportional to the bandwidth of F_1 (second format frequency) [6].

Most pitch detection problems occur at voiced-unvoiced boundaries, where continuity constraints are limited and where pitch period are most likely to be irregular. Figure (1) shows the pitch period of the word /واحد/ and how the period seems to be self-affinity fractal feature. While figure (2) shows the period of melody and how the period seems to be self-similarity fractal feature [7]

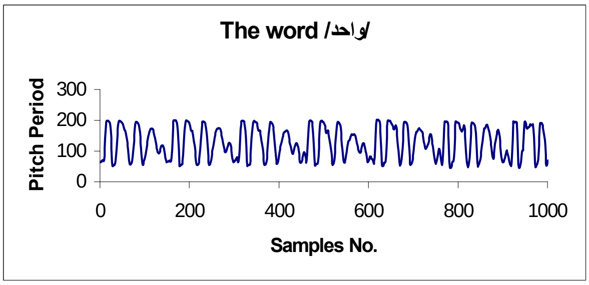

Fig. (1): The pitch period and how it seems to be self-affinity

Fig. (2): The notes of melody and how it seems to be self-similarity

Experimental Results

Practically, it is found that there are two important factors by which can be determine the similarity measure between any two records for the same word pronounced by the same speaker; the ***tone*** of the word and the ***voice loudness***.

The tone is the way of pronouncing a word, it varies from one person to another, especially between the two sexes. This difference is due to prolonging, stressing, or speeding in pronouncing the phoneme of the word. This difference can be exceeded by putting the word within a text to make its pronunciation following, or by practicing its pronunciation in a required way. The voice loudness results to force the air through the larynx cavity, which causes the increase in the vibration of the vocal cords that increases the number of samples per second received by the microphone. Therefore, the distribution of the samples per second appears a dense on the waveform.

These two factors affect the parameters used in the verification process. When the tone and loudness are both fixed, the text can be segmented into words depending on the amplitude of the wave and neglect the unvoiced regions, then store the averaging of extracted verification parameters (Fractal dimension and pitch period), in a code book to be an information of the speakers. It was noticed that the determined fractal dimension shows a signification deviation around its median value. The degree of deviation extends up to extend that the deviated value interferes with the *D*-range of the other words, as shown in figure (3). This reason of using the median filter of the resulting fractal dimension values.

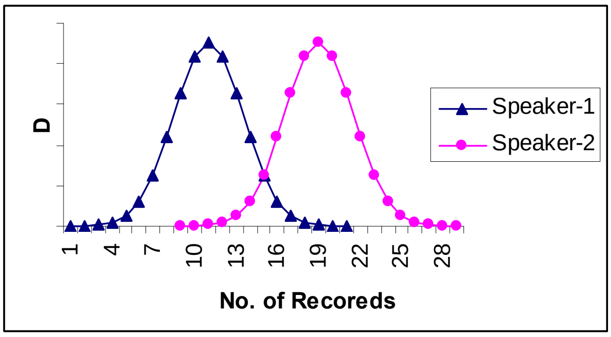

Fig. (3): Fractal dimensions interference

The next stage is the verification. The verification stage is a comparison between the extracted D and PP for unknown speaker with others stored in a code book previously. The stored data follow known speaker, therefore, one can verify the speaker by using the stored data. The verification process by the fractal dimension can be summarized by the following algorithm;

```
Begin
 Read wave file record
 Filter the record
 Find the fractal dimension of the record ($D_o$)
 ND=0:I=0:T=Threshold
 Do
   I=I+1
   If |Di-Do| < T then
     ND=ND+1
     Deference(ND)=|Di-Do|
     Serial(ND)=I
   End if
 Loop Until there is no Di exist in the codebook
 Min=0
 For j=1 to Nd
   If Deference(j)< Min then
     Min= Deference(j)
     K=j
   End if
 Next j
 Ser=Serial(k)
 Identified_D=D(Ser)
 Speaker_Name$=N(Ser)
End
```

By excluding all the fractal dimension values deviated from the median by a difference larger than the standard deviation, the correct identification result between the determined fractal dimension and the code book gave a speaker verification percentage about 70%.

The pitch period was estimated by using a new suggested simple method, which consist of the two following stages;

A- Pitch Extraction. For every sample in speech wave, the average residual (which is the absolute value of the difference between the current sample and N-adjacent neighbors) represent a point in the pitch period.

B- Pitch period detection. The beginning of the pitch period is the beginning of the maximum deviation peak, and its end is the beginning of the next closest similarity peak.

Pitch redundancy, found, differs from speaker to another for same spoken word, and from word to another for the same speaker. But it has same shape for the same word pronounced by the same speaker. Therefore, for identification process, the results of utilizing the pitch period to recognize speakers are well when those speaker are pronouncing the same utterance because of the pitch period is a set of points, so the verification task will be best. The expected better result gave a verification percentage about 90%.

We used G-statistics to study the similarity measure or the behavior of any two resulting curves (two verification parameters). G-statistics estimate the similarity between two distributions to give a number lay within the range between 0 and 1, the low results of G-test indicate that the degree of similarity is high and vice versa. Practically found that the results of order 10^{-3} means the tested pitch periods belong to the same person and those have larger than 10^{-3} indicates dissimilar persons.

By several tests for same and different speakers who are pronounce the same utterance it is notice that both the two parameters (Fractal Dimension and Pitch period) have the nomination to be a good speaker verification parameters, the agreement of them gave a verification percentage of about 85% with equal contribution weight to pronouncement the correct decision.

Figure (3) shows the pitch period of four similar words pronounced by same speaker, figure (4) shows the pitch period of same words pronounced by different speaker, figure (5) shows two pitch periods for different female speakers pronounce same spoken word, and figure (6) presents the pitch period of two different child speakers pronounce same spoken word.

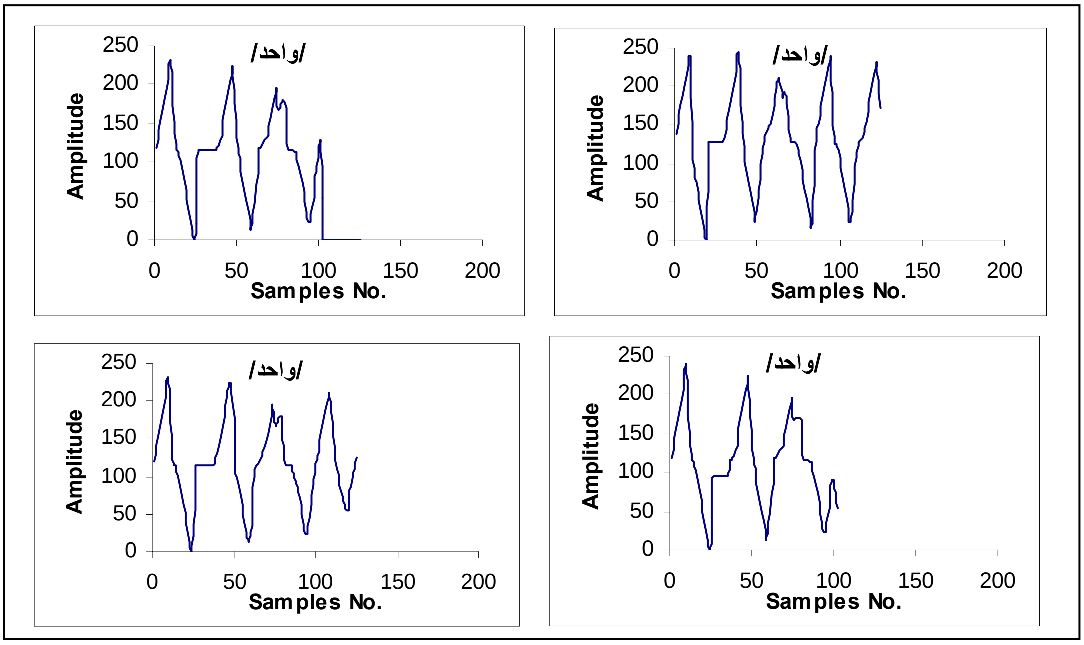

Fig. (4): The pitch period for same male speakers

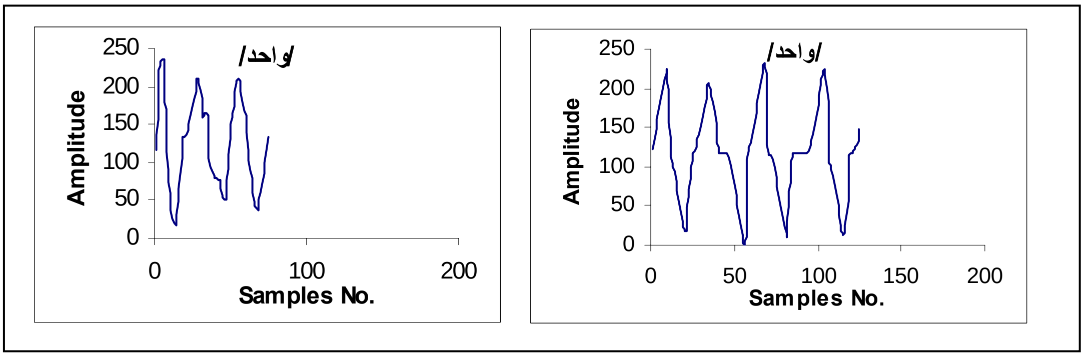

Fig. (5): The pitch period for different male speakers

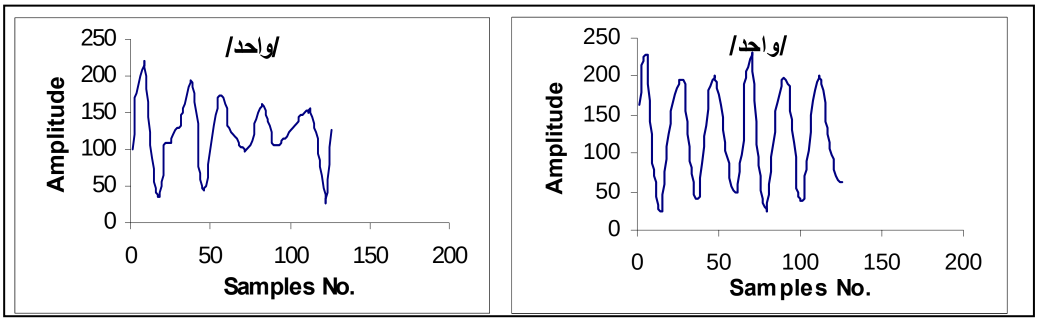

Fig. (6): The pitch period for different female speakers

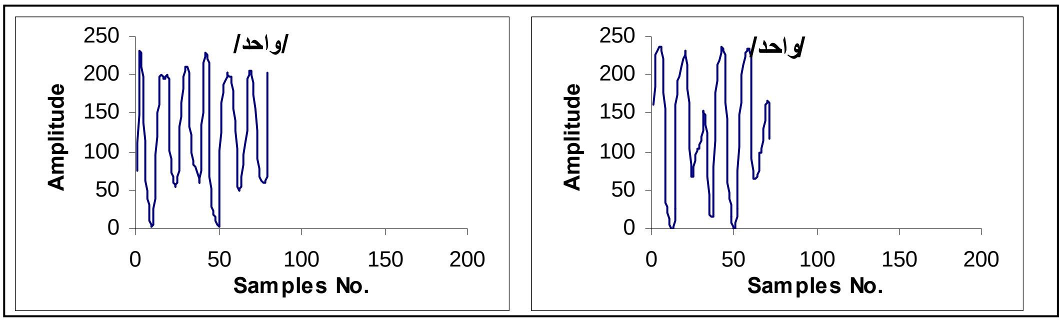

Fig. (7): The pitch period for different children speakers

Conclusions

Practically, it was found that the fractal dimension is inversely proportional with both the speech signal amplitude and the sampling rate, and directly proportional with sound frequency of the record. The results indicated that the fractal dimension is very sensitive even to the very tiny differences that occurring tone of the word during the pronunciation.

In fact, the value of the fractal dimension for the loud voice is greater than that for low voice. Also, its value for female and children is greater than that of the male.

5. References

1. Bransley M. F., **1993**, "*Fractal Every Where*", Academic press Professional Copyright.
2. Keller J. M., Crownover R. M. and Chen R. Y., "*Characteristics of Natural Scenes Related to the Fractal Dimension*", IEEE, Transaction on Pattern Analysis and machine Intelligence, vol. PAMI-9, No. 5, pp. 621-627, September.
3. Desachy J., **1994**, "*Image and Signal Processing for Remote Sensing*", Proceeding Europto Series, vol. 2309, pp. 124-129.
4. Iodic A., Migliaccio M. and Riccio D., **1994**, "*SAR Imagery Classification: The Fractal Approach*", Proceeding Europto Series, vol. 2315, pp. 539-551.
5. Lynn P. A., **1997**, "*Digital Signal processing with Computer Application*", Springer-Verlag New York inc.
6. Mohammed S.M., **2001**, "*Quantitative Analysis on Using Fractal for Sound Signal Processing*", Thesis, Saddam University.
7. Edlr B. and Purnhagen H., **2000**, "*Parametric Audio Coding*", Proceeding of ICCT 2000, August 21-25 2000,Goldton, Beijing,China.