THE FRACTAL BEHAVIOUR OF UNVOICED PLOSIVES: A MEANS FOR CLASSIFICATION

Anastasios Delopoulos and Maria Rangoussi

Computer Science Division,
Department of Electrical Engineering,
National Technical University of Athens,
Athens GR-15780, GREECE
Tel.: +30 1 772 24 91, Fax: +30 1 772 24 92
e-mail: {adelo, maria}@image.ntua.gr

ABSTRACT

Investigation of the fractal behaviour of unvoiced plosive consonants leads to interesting observations towards their classification. Experimental evidence of the fractal nature of the speech signals themselves, as well as of their derivatives and cumulative sums prompt the use of the associated fractal dimensions to form a discriminative feature set. The obtained feature set is compact in representation and easy to compute. At the same time, the discriminating capability of this feature set is seen to be promising even for speech signals sampled at 8KHz.

1. INTRODUCTION

Unvoiced plosives or stops (/k/, /p/ and /t/) correspond to nonstationary, irregular and aperiodic random processes which, unlike voiced sounds, can not be modelled by linear ARMA-type models. Figure 1 depicts a typical example of each one of these sounds.

This inherent irregularity makes modelling and recognition of unvoiced plosives a challenging task. Standard speech recognition methods which attempt to perform classification via context-dependent approaches (see e.g., [2], [3], [7]) have limited applicability, as they require that the unknown plosive lie in a certain context (e.g., triphone models).

Previous attempts for their context-independent characterization include non-parametric, time - frequency representations. In [8], for example, features extracted from the Wigner distribution are employed in appreciation of the non-stationary nature of unvoiced stops.

In this work, the fractal nature of unvoiced plosives is investigated. It is experimentally verified that all three unvoiced plosives as well as their time integrals and derivatives exhibit a fractal behaviour, with significant self-similarity in an extended range of scales. Moreover, this behaviour is consistent within the same sound, while it is diversified across sounds.

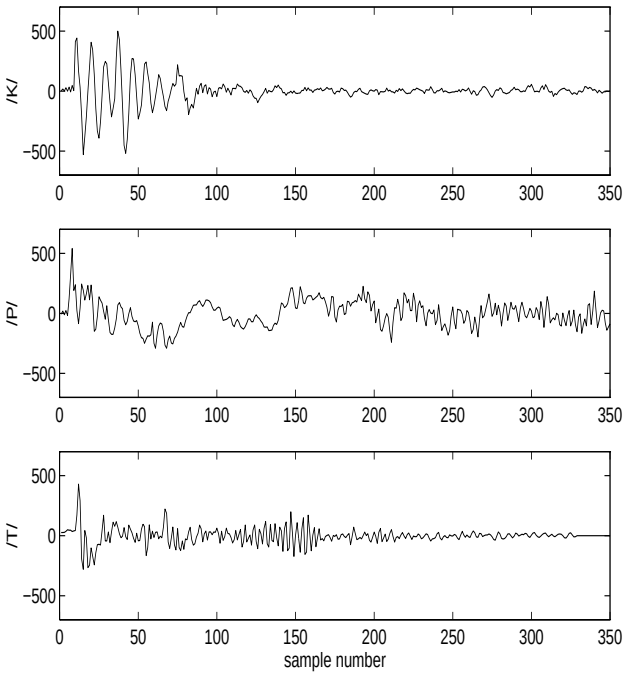

Figure 1: An example of unvoiced plosive consonant signals /k/, /p/ and /t/.

Based on these observations, the aim of the present work is to propose a finite set of parameters which are discriminative enough to qualify for a classification feature vector.

The proposed use of fractal-based features for classification (i) offers a computationally attractive alternative to non-parametric time-frequency based methods, (ii) is parametrized by a much smaller set of parameters - an advantage for the classification step itself. The resulting classification scheme can either be considered as a stand-alone tool or be incorporated in a standard context-based classification method to improve its performance.

In the following, notation related to fractals and fractal dimension is established in context of speech

signals in Section 2 while in Section 3 are given the proposed feature set and its properties along with examples for their computation from real speech data.

2. BACKGROUND ON FRACTAL MEASURES

Fractals have been used in signal processing thanks to their ability to model self-similarities in the domain of signals and/or their statistics (moments) observed over a range of different scales (resolutions), ([1]). Fractals also provide a means for modelling long-term dependencies in the signal, corresponding to long tails of the moment statistics. In particular, the use of fractals in speech processing has been treated in [5].

Fractal dimension, a measure of the fragmentation of a fractal signal, has been defined in various forms, ([1]). Several practical methods for its estimation have been proposed. In the present work we adopt the Minkowski-Bouligand dimension definition D_M, and use the *morphological covering* method proposed in ([6]) for its estimation,

$$D_M = \lim_{\epsilon \to 0} \left(2 - \frac{log[A(\epsilon)]}{log(\epsilon)} \right), \qquad (1)$$

where $A(\epsilon)$ represents the area between the *dilation* and the *erosion* of the signal graph, obtained using a structure element of size proportional to ϵ. In practice, for discrete-time signals, the limiting structure element is of radius 1, corresponding to either a 3×3 rectangle, or a 5-point rhombus or a 3-point horizontal segment. A computationally efficient 1-D procedure for the implementation of the morphological covering of 1-D signals has been proposed in ([6]).

In the neighbourhood of zero, eq. (1) can take the approximate form

$$log \frac{A(\epsilon)}{\epsilon^2} = D_M \; log \left(\frac{1}{\epsilon}\right) + constant, \qquad (2)$$

which represents a line of slope D_M in the $(log \frac{A(\epsilon)}{\epsilon^2}, log \left(\frac{1}{\epsilon}\right))$ plane. For one-dimensional signals, a value of $1 < D_M < 2$ signifies a fractal, while non-fractal signals have $D_M = 1$.

For signals that are ideal fractals, equation (2) holds true not only for ϵ lying in the neighbourhood of zero, but for any ϵ. However, real life signals are rarely exhibiting such a perfect behaviour. Therefore, for such signals it is meaningful to examine equation (2) in a *local* manner. This is equivalent to fitting straight line segments locally, after partitioning the range of $\log \left(\frac{1}{\epsilon}\right)$ under examination into a set of successive disjoint intervals. This procedure produces a sequence of different slopes which hereafter we call *Local Fractal Dimension (LFD),* after [6].

3. FRACTAL DIMENSION OF UNVOICED PLOSIVES

In order to investigate the fractal behaviour of unvoiced plosives, the fractal dimension of

(i) the speech signals themselves, $s(n)$,

(ii) their running averages (cumulative sums), $c(n) = \sum_{i=0}^{n} s(i)$ and

(iii) their discrete derivatives (increments), $d(n) = s(n) - s(n-1)$,

has been computed as the slope of the line in equation (2). The slope has been computed for a set of 100 decreasing scales ϵ. The motivation for considering the fractal properties of $c(n)$ and $d(n)$ comes from the case of the fractional Brownian motion, where it is the increments rather than the process itself which exhibit fractal behaviour, [4].

For the purposes of this study a pool of 50 unvoiced plosive sounds (20 /k/, 10 /p/, 20 /t/) were extracted from TIMIT speech database, at an 8 KHz A/D sampling rate, without any constrain as to the context, the speaker identity or the speaker sex.

Figures 2.a, 2.b and 2.c show the fractal dimension estimates for $s(n), c(n), d(n)$, respectively. All three plots are indexed by the signal number, (for /k/ $(1, \ldots, 20)$, connected by solid line, for /p/ $(1, \ldots, 10)$, connected by dashed line and for /t/ $(1, \ldots, 20)$, connected by dotted line).

The main observations made from these figures are that

1. fractal dimension D_M is significantly greater than 1 for all the cases examined, which verifies the fractal nature of $s(n)$, $c(n)$ and $d(n)$;

2. the behaviour of sound /p/ in the running average domain (Figure 2.b, 10 middle values) is systematically different than that of /k/ and /t/. This makes the fractal dimension D_M of the running average $c(n)$ a candidate feature for discrimination among /p/ and the other two sounds;

3. /k/ and /t/ can not be discriminated on the basis of D_M alone.

In view of the last observation, the local fractal dimension, LFD, has been examined for the same signal cases of $s(n)$, $c(n)$ and $d(n)$. For the purposes of this experiment, the $log \left(\frac{1}{\epsilon}\right)$ horizontal axis was partitioned in $L = 20$ non-overlapping intervals of equal length, covering $0 < \epsilon < 1$. The values of the LFD sequences have been used to form vectors $\mathbf{l}_s$, $\mathbf{l}_c$ and $\mathbf{l}_d$ of dimension $L \times 1$, for $s(n)$, $c(n)$ and $d(n)$, respectively.

From the study of the obtained LFDs it was concluded that increased separability of the three sounds is obtained on the basis of those LFD coefficients corresponding to the finest scales (resolutions). Figures

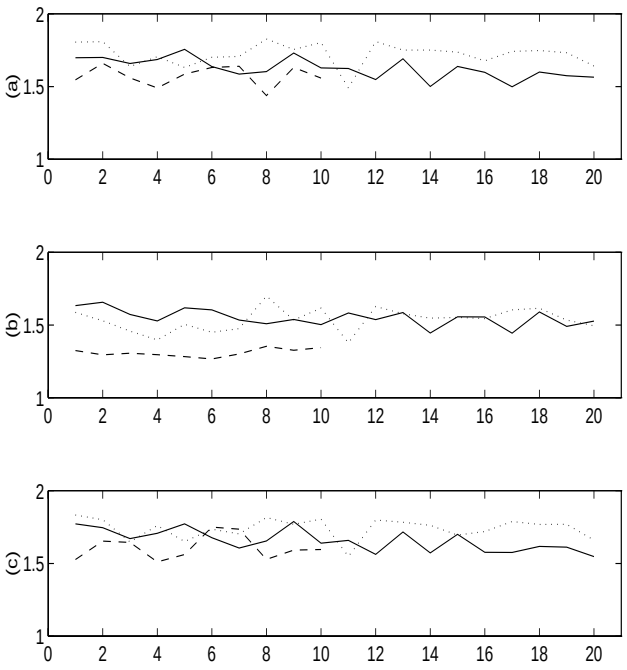

Figure 2: Fractal dimensions of signals (a) $s(n)$, (b) $c(n)$ and (c) $d(n)$. Each subplot shows the fractal dimension D_M of (i) the 20 /k/ signals (solid line), (ii) the 10 /p/ signals (dashed line) and (iii) the 20 /t/ signals (dotted line).

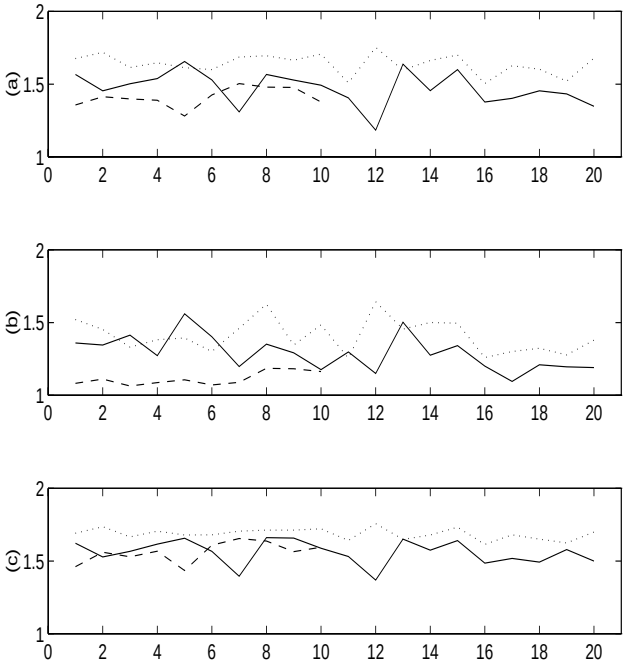

Figure 3: Local fractal dimensions of signals (a) $s(n)$, (b) $c(n)$ and (c) $d(n)$ at the finest scale $L = 20$. Each subplot shows the LFD $\mathbf{l}(L)$ of (i) the 20 /k/ signals (solid line), (ii) the 10 /p/ signals (dashed line) and (iii) the 20 /t/ signals (dotted line).

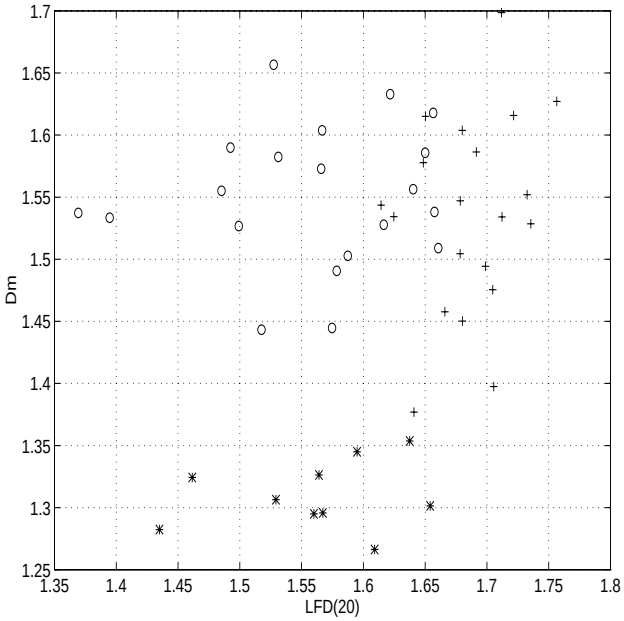

Figure 4: The spread of the proposed two-dimensional feature vector v in the feature space. /k/, /p/ and /t/ are denoted by circle, star and cross signs, respectively. Even a linear classifier would achieve high classification scores.

3.a, 3.b and 3.c show the LFD values obtained in the finest of the $L = 20$ scales examined, for $s(n)$, $c(n)$ and $d(n)$, respectively. The horizontal axis is indexed by the signal number, as in Figure 2. As it can be seen in Figure 3, the derivative $d(n)$ is more discriminative with respect to the /k/ and /t/ behaviour.

Consequently, for classification of the three sounds the proposed feature vector includes the features

$$v = [D_{M,c}, \mathbf{l}_d(L)]. \tag{3}$$

Figure 4 shows the feature values of v computed for the set of 50 samples from the TIMIT speech database, as mentioned above. Features obtained from /k/, /p/ and /t/ are denoted by circle, star and cross signs, respectively. From this figure it can be seen that the proposed feature vector v yields satisfactory class separation. Moreover, class separability is expected to improve if signals of a higher sampling rate are used.

Finally, the consequent classification step can be carried out using any standard classifier, such as LVQ, although even simpler linear classifiers might be adequate, as it can be inferred based on Figure 4.

4. CONCLUSIONS - FURTHER RESEARCH

The fractal behaviour of unvoiced plosive consonants is experimentally verified in the present work. Moreover, the same behaviour is exhibited by their discrete-

time derivative and cumulative sum signals. Although the time plots of unvoiced stops do not allow even for their crude classification by inspection, since they do not exhibit a systematic similarity, their fragmentation and the associated fractal consistently assume ranges of values that allow for their characterization. The proposed feature set is compact in representation and promising for utilization in a context- and speaker-independent speech recognition task.

More extensive experimentation is necessary, however, in order to obtain statistically significant classification scores. On the other hand, given the fact that speech signals are the output of the human speech generation mechanism, it is interesting to conduct further research in order to reveal the physiological factors that are responsible for this behaviour.

5. REFERENCES

[1] M. F. Barnsley, "Fractals Everywhere," New York, Academic Press, 1988.

[2] K-F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Trans. on ASSP*,, vol. 38, no. 4, pp. 599-609, April 1990.

[3] K-F. Lee, H. W. Hon, R. Reddy, "An Overview of the SPHINX Speech Recognition System," *IEEE Trans. on ASSP*,, vol. 38, no. 1, pp. 35-45, Jan. 1990.

[4] B. Mandelbrot, J. W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Rev.,* vol. 10, pp. 422-437, Oct. 1968.

[5] P. Maragos, "Fractal aspects of speech signals: Dimension and Interpolation," *Proc. IEEE ICASSP'91,* Toronto, Canada, May 1991.

[6] P. Maragos, F.-K. Sun, "Measuring the Fractal Dimension of Signals: Morphological Covers and Iterative Optimization," *IEEE Trans. on Signal Processing,* vol. 41, no. 1, pp. 108-121, Jan. 1993.

[7] K. S. Nathan, H. F. Silverman, "Time-Varying Feature Selection and Classification of Unvoiced Stop Consonants," *IEEE Trans. on Speech and Audio,* vol. 2, no. 3, pp. 395-405, July 1994.

[8] M. Rangoussi, A. Delopoulos, "Recognition of Unvoiced Stops from their Time-Frequency Representation," *Proc. IEEE ICASSP'95,* Detroit, USA, May 1995.