A COMPARISON OF WAVEFORM FRACTAL DIMENSION TECHNIQUES FOR VOICE PATHOLOGY CLASSIFICATION

Pallavi N. Baljekar

Dept. of Electronics and Communication
Manipal Institute of Technology (MIT),
Manipal, Karnataka, India.
pallavi.baljekar@learner.manipal.edu

Hemant A. Patil

Dhirubhai Ambani Institute of Information and
Communication Technology (DA-IICT),
Gandhinagar, Gujarat, India.
hemant_patil@daiict.ac.in

ABSTRACT

In this paper, an attempt is made to compare and analyze the various waveform fractal dimension techniques for voice pathology classification. Three methods of estimating the fractal dimension directly from the time-domain waveform have been compared. The methods used are Katz algorithm, Higuchi algorithm and the Hurst exponent calculated using the rescaled range (R/S) analysis. Furthermore, the effects of the window size, the base waveform used and score-level fusion with Mel frequency cepstral coefficients (MFCC) has also been evaluated. The features have been extracted from two different base waveforms, the speech signal and the Teager energy operator (TEO) phase of the speech signal. Experiments have been carried out on a subset of the Massachusetts Eye and Ear Infirmary (MEEI) database and classifier used is a 2^{nd} order polynomial classifier. A classification accuracy of 97.54 % was achieved on score-level fusion, an increase in performance by about 2 % as compared to MFCC alone.

Index Terms— Fractal dimension, Voice Pathology, Hurst exponent, Higuchi algorithm, Polynomial classifier.

1. INTRODUCTION

Pathologically affected voices are characterized by their increased nonlinearity and turbulent nature. This study investigates the effectiveness of using a nonlinear feature, *viz.,* the waveform fractal dimension as a correlate of the nonlinearity and turbulent nature of a patient's speech signal. The main aim in investigating such features is to develop a robust and accurate method for detecting pathologies affecting the vocal folds and the vocal tract in a convenient, accurate and noninvasive manner. The voice disorders can be broadly classified as functional and organic disorders. Organic disorders are caused because of some physical malfunction in the voice production mechanism, while functional disorders results from misuse of vocal production mechanism [1]. The presence of these pathologies modifies the speech signal by introducing noise transients, due to asymmetric vibration and incomplete closure of vocal folds. Hence, the perceived pathologically affected voice sounds hoarse and breathy.

The Teager energy operator (TEO), proposed by the Teagers [2], characterizes the nonlinearity in the vocal production mechanism by accounting for the nonlinear sources of voice production mechanism, *viz.,* the vortices caused due to the turbulent and nonlaminar nature of the airflow. TEO phase is obtained as the cosine of the analytic phase of this TEO profile [3]. Fig. 1 illustrates the process of obtaining the analytic phase of the TEO profile of a speech signal *x(n)*. The speech signal and the corresponding TEO phase profiles for a normal speaker and a speaker suffering from vocal folds polyps has been plotted in Fig. 2. The differences in the structure, periodicity and roughness of the waveforms (both speech and TEO phase) can be observed on comparing Panel 1 corresponding to the normal speech and Panel 2, corresponding to the speaker suffering from vocal fold polyps. By comparing the speech signal and TEO phase of the speaker suffering from polyps, it can be seen that the TEO phase tends to amplify the transients present in the speech signals. This may presumably be because of the ability of the TEO phase to capture the instantaneous phase changes in the speech waveform [3]. Since it is known that the fractal dimension quantifies the roughness of a surface or in other words the transient behavior of the waveform [4], in this work, the fractal dimension (FD) extracted from two base waveforms, the speech signal and the TEO phase of speech has been analyzed.

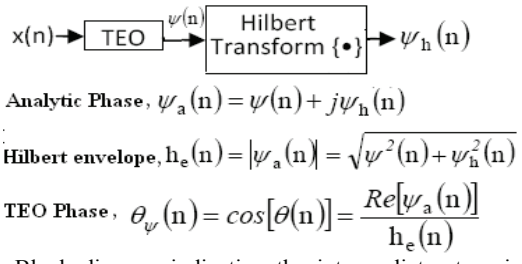

Fig 1. Block diagram indicating the intermediate steps in TEO phase computation.

The FD measures the dimension of a geometric object, which can be a non-integer number. The greater the FD of an object, the greater is the roughness or the irregularity of the surface of the object, which implies a greater presence of transient behavior. This FD can either be measured by assuming the waveform itself as the geometric object or it can be obtained from the phase space, by

measuring the FD of the attractor in the phase space. It has been used previously on various biomedical signals like electroencephalogram (EEG), electrocardiogram (ECG), speech, etc., because of the inherent nonlinearity present in biological control systems [4]-[5]. In the application of detecting pathological voices, methods using FD derived from the phase space have been used [5]. In this paper, we investigate three well known algorithms to compute the FD of a waveform, *viz.*, Katz's algorithm, Higuchi's algorithm and FD obtained using the Hurst exponent, which is calculated using the rescaled range analysis *(R/S)*. These time-domain FD extracting methods are computationally more efficient and much simpler to implement than the phase space methods.

The paper is organized as follows. Section II gives the details of the algorithms used to calculate the FD. Section III describes the methodology and analyses the experimental results obtained. Lastly, Section IV, summarizes our conclusions and lists our future research directions.

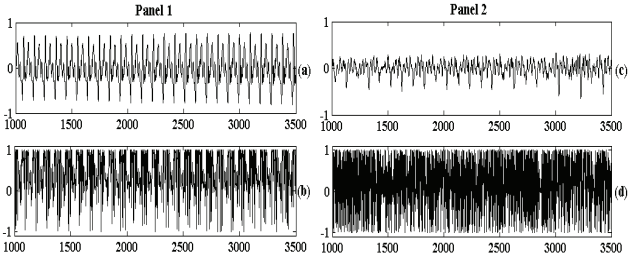

Fig 2. Plots of the speech waveforms for (a) a normal speaker and (c) speaker suffering from vocal folds polyps and corresponding TEO phase for (b) normal (d) polyps.

2. WAVEFORM FRACTAL DIMENSION ALGORITHMS

This section describes the three methods used to estimate the FD and provides a physical interpretation of each of the algorithms.

2.1. Katz Algorithm

This algorithm describes the FD, denoted as *D* as a ratio of the length of the curve *L*, calculated as the sum of the distances between two successive points, divided by the maximum distance *d* of any point in the frame under consideration from the first point. Thus, it can be interpreted as the ratio of the total length of the curve as compared to the straight line corresponding to the maximum linear Euclidean distance from the first point. Hence, it measures the extra length of the curve as compared to the maximum linear distance and thus could be viewed as a measure of the *roughness* of the waveform, since the more rough a curve the greater would be its length as compared to a straight line between the two furthest points on the curve. Moreover, since the distances between successive points would depend on the sampling frequency, a scale factor $\overline{a}$, is used which is the average of the distances between two successive points. The fractal dimension *D* is thus defined as:

$$D = \frac{log_{10}(L/\bar{a})}{log_{10}(d/\bar{a})}, \quad (1)$$

where *L* is the total length of the waveform, *d* is the maximum distance between the first point and any other point on the waveform and $\overline{a}$ is the scale factor.

2.2. Higuchi Algorithm

The FD using Higuchi algorithm is calculated as in [4]. It consists of forming new waveforms by iteratively selecting samples differing in their starting point *m* and their delay factor *k*. We first select a maximum delay factor, say k_{max}. So for every delay factor *k*, where *k* is varied from *1* to k_{max}, we form *k* new time series, x_m^k, where the starting point of the series is defined by *m* and samples at every *k* samples are selected to form the new waveform, i.e.,

$$x_m^k = \{x(m), x(m+k), x(m+2k).x(m+\lfloor (N-m)/k \rfloor k\}, \quad (2)$$

where *m* the starting point for each new waveform is varied from *1* to *k* , *k* is the delay factor between the samples and *N* is the window size. Then the length of each waveform is calculated as the sum of the distances between two consecutive points, i.e.,

$$L(m,k) = \frac{\sum_{i=1}^{\lfloor (N-m)/k \rfloor} |x(m+ik) - x(m+(i-1)k)|(N-1)}{\lfloor (N-m)/k \rfloor k}, \quad (3)$$

where $\lfloor (N-m)/k \rfloor k/(N-1)$ is the normalization factor and *N* is the window length. The lengths for the same delay factor, *k* are then averaged as follows,

$$L(k) = \sum_{m=1}^{k} L(m,k), \quad (4)$$

where *k* is varied from *1* to k_{max} and *L(k)* is the averaged length for a particular delay factor *k*. We would expect that for very smooth and regular waveforms, as the delay factor is increased, the length of the waveform would decrease proportionally since increasing the delay factor between samples, could be viewed as smoothening the waveform, and hence we would expect the length to decrease proportionally. However, if it is an irregular waveform, containing a lot of transients, this decrease in length will be very much, since by neglecting samples in between, we are bypassing many of the transients in between and as a result with increasing *k*, we expect a very steep decrease in *L(k)*. The FD is thus calculated as the slope of the least squares linear best fit to the graph plotted between *ln(L(k))* and *ln(1/k)*. So greater the slope, greater is the number of transients in the waveform and hence greater is its FD.

2.3. Hurst Exponent

The Hurst exponent has mostly been used in analysis of financial time series in order to predict the trend [6]. In this paper, the Hurst exponent is calculated using the rescaled range analysis *(R/S)*. At each point in a region of size *n*, the cumulative deviation upto that point is calculated and stored in a vector *X(t)*. The range R_n for that region is calculated as the difference between the maximum and minimum value of *X(t)*, i.e.,

$$R_n = \{max(X(t)) - min(X(t))\}, \quad (5)$$

where, *n* is the number of samples in the region under consideration. The denominator S_n is the standard deviation of the region of size *n* under consideration, i.e.,

$$S_n = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x(t)-\mu\right)}\ , \qquad (6)$$

where *x(t)* are the samples of the frame of the waveform under consideration and μ is the mean of the samples in that frame . The *R/S* value is then averaged over all the *R/S* values for a given region size. In this way by iteratively dividing the dataset by a factor of 2, till it reaches a very small region size, (in our experiments we considered the smallest region of 16 samples) the *R/S* values were obtained and a graph was plotted between $log_2(R/S)$ value *vs*. the $log_2(n)$. The slope of this line is the estimate of the Hurst exponent (*H*). An *H* value of 0.5 indicates a total random walk. A relatively larger value of $0.5 < H < 1$, indicates persistent behavior, i.e., if a waveform is in an increasing trend it will remain in an increasing trend or vice-versa. Whereas, a lesser value, i.e., $0 < H < 0.5$ indicates anti-persistent behavior, i.e., a waveform in an increasing trend will most likely decrease or vice-versa. What this implies is that for a smooth and regular waveform, we expect there to be a proportional increase in *R/S* value with an increase in *n* to give a value of *H* as very close to 1. This is because, we expect that for smooth waveforms, as the region size *n* increases, the range *R* also increases since there would be regions in an up trend where the cumulative deviation would be adding up to a large positive number and regions in a down trend where the cumulative deviation would be adding up to a large negative number thus yielding a large *R* value as compared to the total deviation *S* of the region. This may not be true in the case of irregular waveforms since the *R/S* value may not necessarily increase proportionally with the increase in *n* due to the increased irregularities yielding a lower *R/S* value. The fractal dimension is then calculated from the Hurst exponent using the following relation [6]:

$$D = 2 - H\ . \qquad (7)$$

Thus, it can be said that the Hurst exponent measures the short-term predictability of a signal.

3. EXPERIMENTS

This section gives the details of the feature extraction method and the database and analyses the results obtained.

3.1. Data and Methods

In this work, for each algorithm, the effect of three basic criteria were explored, namely, window size, base waveform and complementary information provided on score-level fusion with state-of-the-art Mel frequency cepstral coefficients (MFCC). The database used for the experiments was the commercially available Massachusetts Eye and Ear Infirmary (MEEI) database [7]. In this study, a subset of this database consisting of 53 normal speakers forming the control group and 173 speakers suffering from various pathologies according to the speech corpus design given in [8] was considered for classification. All the samples were downsampled to 25 kHz sampling frequency. For feature extraction, the base waveform, i.e., speech (Sp) or TEO phase (TP), was first blocked into frames of *N* samples with 50 % overlap. In this work, three values of *N* were considered, *N*=256, 512 and 1024 samples. Each frame was then multiplied with a Hamming window of length *N* and the short-term FD was extracted per frame. This FD vector was then fed to a *discriminatively trained* 2^{nd} order polynomial classifier [9] to generate the true and false scores, using which the detection error tradeoff (DET) curves were plotted. The equal error rate (EER) obtained from these DET curves and the accuracy (Acc %) were used as performance measures. These values have been shown in Table 1. Score-level fusion was carried out with MFCC and FD. For MFCC computation, each frame was blocked into frames of 256 samples with 50 % overlap, and MFCC computation was carried out according to [10], with 12 MFCC coefficients extracted per frame. These features were then given to a polynomial classifier to generate the true and false scores. The scores obtained for MFCC were then fused in equal weights (i.e., for *w*=0.5), with the scores of the best performing feature in each algorithm (highlighted in Table 1) according to (8).

$$Y_f = wY_{MFCC} + (1-w)Y_x \qquad (8)$$

Here, *x* represents each of the best performing FD features highlighted in Table 1 and Y_{MFCC}, Y_x and Y_f are matching scores for MFCC, FD and their score-level fusion, respectively. These fusion results have been shown in Table 2. A four-fold cross-validation scheme was used repeated 12 times, giving a total of 48 trials so as to produce a smooth DET curve and make the classification independent of the training and testing set. For each trial, 75 % of the samples from each class (selected randomly) were used for training and the rest 25 % were used for testing.

Table 1. EER (%) and accuracy (Acc (%)) values obtained using various waveform fractal dimension algorithms for varying window size and base waveform.

Algorithm/ Feature	Base Waveform	Frame Size	Acc (%)	EER (%)
MFCC	**Speech**	**256**	**95.65**	**4.35**
Katz	Speech	1024	66.22	36.27
Katz	Speech	512`	71.65	32.63
Katz	Speech	256	68.78	34.11
Katz	TEO Phase	1024	82.18	17.82
Katz	TEO Phase	512	82.51	17.48
Katz	**TEO Phase**	**256**	**82.18**	**17.82**
Higuchi	Speech	1024	59.38	40.63
Higuchi	Speech	512	59.45	40.55
Higuchi	Speech	256	59.60	40.40
Higuchi	TEO Phase	1024	80.92	19.08
Higuchi	TEO Phase	512	81.40	18.60
Higuchi	**TEO Phase**	**256**	**81.62**	**18.38**
Hurst Exp	**Speech**	**1024**	**87.72**	**12.28**
Hurst Exp	Speech	512	87.46	12.54
Hurst Exp	Speech	256	85.52	14.47
Hurst Exp	TEO Phase	1024	75.26	24.74
Hurst Exp	TEO Phase	512	73.81	26.19
Hurst Exp	TEO Phase	256	75.52	24.48

Table 2. EER(%) and accuracy (Acc(%) values of score-level fusion of the best performing features with MFCC.

Feature	Base signal	Frame Size	Acc (%)	EER (%)
FD (Hurst) + MFCC	Sp	1024	97.54	2.45
FD (Higuchi) + MFCC	TP	256	96.39	3.61
FD (Katz) +MFCC	TP	512	95.91	4.09

3.2. Result Analysis

The results of the experiments carried out have been shown in Tables 1 and 2. The observations from these tables are as follows.
- *Base waveforms*: It can be observed that two out of the three algorithms considered give superior performance for FD extracted from TEO phase as compared to the speech waveform across all window sizes. This maybe due to the fact that the different algorithms consider different parameters as a measure of the FD. For instance, the Katz and Higuchi algorithms, consider the extra length of the curve as a measure of the FD, while the FD obtained using the Hurst exponent considers the range of the cumulative deviations in a given waveform as a measure of the FD.
-*Window size*: As such, the difference in the accuracy or the EER with a change in window size is not very significant. In the Katz algorithm, it can be observed from Table 1 that for both the base waveforms, the window size of 512 samples seems to be the most optimum. While in case of the Higuchi algorithm, we see that the accuracy and the window size are inversely proportional to each other, i.e., as the window size is decreased, the accuracy increases. Finally, considering the Hurst exponent method, we see that for the speech waveform as the base waveform, there is a direct relationship between the accuracy and window size, but for the TEO phase as the base waveform, there is no trend as such, since a window size of 1024 and 256 give comparable results. However, as such we can conclude that for Hurst exponent method, longer windows generally tend to give a higher accuracy.

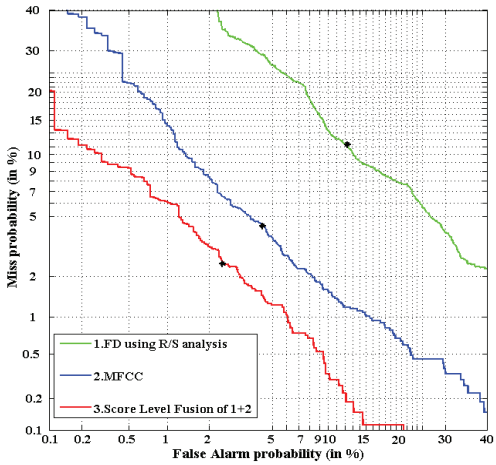

Fig 3. DET plot showing the complementary nature of FD extracted using the Hurst exponent from speech signal for N=1024.

-*Score-level Fusion:* The scores of the best feature from each algorithm were fused with the scores of MFCC to investigate whether the FD gave some complementary information. As can be seen from Table 2, the score-level fusion of the short-term FD with MFCC does give some complementary information in all cases since in each case the EER is lesser than that of MFCC alone which is 4.35% and the accuracy is greater than that of MFCC alone *viz.,* 95.65%. Fig. 3 shows the DET plot for MFCC, FD extracted using Hurst exponent and their score-level fusion. It can be seen that in the DET plot shown in Fig. 3, the score-level fusion performs significantly better at *all* points on the DET curve and it can be observed that there is a significant decrease in the EER by almost 1.9% and an increase in accuracy by almost 2.5% as compared to that of MFCC alone. Thus, we can conclude that the FD does provide certain amount of *complementary* information and can be used to increase the classification accuracy for voice pathology classification.

4. SUMMARY AND CONCLUSIONS

In this paper, we investigated three algorithms for computing the FD directly from the waveform in the time-domain. It was observed that, the performance of the base waveform depends on the quantity that the algorithm considers to be a measure of the FD. As per our results, it was concluded that if this quantity is the length of the curve, the TEO phase performs better, while if it is the cumulative deviation, the speech waveform performs better. Furthermore, it was observed that the optimum window size depends upon the FD extracting algorithm. On score-level fusion with state-of-the-art MFCC feature-set, we found that the FD did provide some complementary information and decreased the EER by almost 2 %. Thus, we can conclude that FD derived directly from the waveform does increase the accuracy of the system when fused with MFCC, without much increase in the computational complexity. The main limitation of using the Katz algorithm is its dependence on the sampling frequency. Hence, in our future studies, we would like to study the effect of the sampling frequency in the FD computation. In addition, in this work for the Higuchi algorithm, we used a fixed commonly used value for the maximum delay factor k_{max} = 8, hence, we would like to investigate techniques to select an optimum k_{max} to improve the classification performance.

5. RFERENCES

[1] P.L. Dhingra, *Diseases of ear, nose and throat*, 3rd ed., Elsevier, New Delhi, 2004.

[2] H. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *In Speech Production and Speech Modelling* , W.J. J. Peters, Ed. McGraw-Hill, New York, pp. 241-261,1990

[3] H. A. Patil and K. Parhi "Development of TEO phase for speaker recognition," *in Proc. of International Conference on Signal Processing and Communications*, Bangalore, 2010, pp. 1-5.

[4] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications,* , vol.48, no.2, pp.177-183, Feb 2001.

[5] J. D. Arias-Londono, J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, and G. Castellanos-Domınguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and Mel-cepstral coefficients," *IEEE Trans. on Biomedical Engg*., vol. 58,no. 2, pp.370-379, Feb 2011.

[6] B. Qian and K. Rasheed. "Hurst exponent and financial market predictability," *IASTED conference on Financial Engineering and Applications*, 2004, pp. 203- 209.

[7] Kay Elemetrics Corp, Disordered Voice Database Model 4337, Ver. 1.03, Massachusetts Eye & Ear Infirmary Voice & Speech Lab, 2002.

[8] V. Parsa and D.G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Language, Hearing Res.,* vol. 43, no.2, pp. 469–485, 2000.

[9] W. M. Campbell, K. T. Assaleh and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Trans. on Speech and Audio Processing*, vol.10, no.4, pp. 205-212, May 2002.

[10] S.B. Davis, and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 28, no.4, pp 357—366, 1980.