EFFECTIVENESS OF MULTISCALE FRACTAL DIMENSION FOR IMPROVEMENT OF FRAME CLASSIFICATION RATE

Mohammadi Zaki, Nirmesh J. Shah and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar - 382007, India

Email: {mohammadi_zaki, nirmesh_shah, hemant_patil}@daiict.ac.in

ABSTRACT

We propose to use multiscale fractal dimension (FD)-based features for phoneme classification task at frame-level. During speech production, turbulence is created and hence vortices (generated due to presence of separated airflow) may travel along the vocal tract and excite vocal tract resonators. This *turbulence* and in effect, the embedded features of different phoneme classes, can be captured by *invariant* property of multiscale FD. To capture complementary information, feature-level fusion of proposed feature with state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) is attempted and found to be effective. In particular, single-hidden layer neural nets were trained to compute the frame classification rate. Proposed feature was able to reduce the error rate by over *1.6* % from MFCC features on *TIMIT* database. This is supported by significant reduction in % *EER* (*i.e.*, *0.327* % to *4.795* %)[1].

Index Terms— fractal dimension, multiscale analysis, phoneme-based frame classification, nonlinearity.

1. INTRODUCTION

Fractal dimension (FD) has been used and applied in various areas of engineering wherever it is essential to investigate the amount of irregularity or ruggedness of the object under consideration. Practically, every object that we encounter in real life is a fractal, as there is no perfectly circular or rectangular or of any other such artificially created object occurring naturally [1]. The FD has been successfully applied to many one-dimensional (i.e., *1-D*) signals (called time series) to measure their *chaoticity* [2].

Inherent property of FD is expected to quantify the amount of nonlinearity in the underlying speech production system. These nonlinearities are explained by the aeroacoustic model of the vocal tract [3], which takes into account the non-acoustic fluid motion in the vocal tract as a secondary source of sound, which is usually neglected by the linear *source-filter* theory. According to Sinder [4], these non-acoustic fluid motion influences the production of wide classes of phonemes (e.g., vowels, voiced and unvoiced plosives, fricatives, etc). However, the amount of non-linearity producing turbulence varies for the different classes, which we attempt to quantify by using FD [5]. Many studies have tried to incorporate such features of different sounds into their pattern recognition methods. However, most of such techniques are used for classification between pathological *vs.* non-pathological voices [6-7], speaker identification [8] and other specific experiments [9-10]. However, it was after studies reported in [5, 11-13, 2] that FD was formally used for recognition of isolated phonemes. Recently, *1-D* FD had been used for ASR task which report improvement in accuracy [14].

This paper is an extension of our previous work for the entire TIMIT database [15]. The spectral transitivity property of multiscale FD is experimented which is carried out by using a *39-D* feature vector consisting of MFCC and FD features. The delta (Δ) and delta-delta (ΔΔ) features of MFCCs are known to capture spectral transitivity in the speech signal. The FD is known to capture the dynamics of the system [10], and will be shown to be a characteristic feature of the different phonemes. In addition, the paper emphasizes on the *multifractal* property of speech signal which is explicitly captured by the *multiscale fractal dimension* (MFD). Here, we investigate capabilities of MFD towards these aspects, by training single layer *feed-forward* neural network which classifies the given input frame into one of the *61* phoneme classes of TIMIT.

Organization of the paper is as follows. Section 2 gives brief connection between speech and FD. Section 3 explains the relevance of MFD in phoneme-based frame recognition task. In Section 4, incorporation of MFD based features with MFCC to capture the nonlinearity in speech frame is explained, so that it can be used in frame classification task. Section 5 presents summary and conclusions.

2. SPEECH AND FRACTAL DIMENSION

There have been many efforts for characterizing the nonlinearities in non-acoustic fluid motion in the vocal tract. According to conservation of momentum, the Navier-Stokes equation for speech production is given by [6], [16].

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), New Delhi, India for sponsoring the consortium project, *viz.,* "Development of TTS for Indian Languages". They also thank the authorities of DA-IICT for supporting this research.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

$$\rho\left(\frac{\partial \mathrm{u}}{\partial t}+\mathrm{u}.\nabla \mathrm{u}\right)=-\nabla p+\mu\nabla^2\mathrm{u}, \quad (1)$$

where ρ is the air density, p is the air pressure, u is the vector air particle velocity and 'μ' is the air-viscosity coefficient, assuming negligible flow compressibility. The Reynolds number (*Re*) is a characterizing parameter for measuring the aerodynamics. It has been shown in [17], [18] that speech production consists of dynamics which have very high value of *Re*, thereby generating turbulent airflow [19]. The turbulent nature of the airflow is caused because of the existence of *eddies* at multiple scales. In addition, by Kolmogorov's law, velocity field can be modeled as a process $V(x)$ whose increments has a variance given by [20]:

$$E\left\{\left|V\left(x+k\right)-V\left(x\right)\right|^2\right\}\propto r^{2/3}k^{-5/3}, \quad (2)$$

where k is the wave number in a finite non-zero range and r is the energy-dissipation rate and $E(k, r)$ is the velocity-wave number spectrum, i.e., Fourier transform of spatial correlations [20]. It is to be noted that r varies with the spatial location x. This essentially states that the energy of the signal varies along with its scale. This difference in energy between various scales is what gives rise to coherent structures such as *vortices* producing turbulence. Hence, it is essential to understand the multiscale characteristics of speech signal because of its inherent turbulence [5], [21].

Multiscale fractal dimension. FD defined by Mandelbrot is given by [1],

$$FD=\lim_{\varepsilon\to 0}\frac{\log\left(N(\varepsilon)\right)}{\log\left(1/\varepsilon\right)}, \quad (3)$$

where $N(\varepsilon)$ is the number of compact planar shapes of size ε required to cover the fractal objects under consideration. However, different speech segments have different levels of turbulence. Hence, *ruggedness* in *1-D* speech signal would be different for different phonemes. Therefore, we need to measure FD at various scales of the covering object. The morphological FD is an efficient method for *multiscale* estimation of FD. The idea behind the algorithm is to measure the area covered by the object at various scales,

$$A_G(\varepsilon)=area(S\oplus\varepsilon G), \quad (4)$$

where S is the object under consideration εG is the ε-scaled covering element and $\oplus$ is the nonlinear morphological *dilation* operation. The fractal dimension, *FD,* is defined as,

$$FD=2-\lim_{\varepsilon\to 0}\frac{\log(A_G(\varepsilon))}{\log(\varepsilon)}. \quad (5)$$

Eq. (4) and eq. (5) represent morphological FD algorithm for continuous case. The discrete version of the above algorithm is presented in [21]. These calculations are for value of $\varepsilon=1$. It has been proven that in order to move to next higher scale of ε, we iterate the process [21]. If we repeatedly perform this for say $\varepsilon=1, 2,\ldots, N$ times then we can achieve FDs with covering elements of size $1/f_s, 2/f_s,\ldots,N/f_s$ seconds, which is known as *multiscale FD*.

3. RELEVANCE OF FD FOR FRAME CLASSIFICATION TASK

The multifractal nature of speech signals has been proved in [22-23]. The singularity spectrum, is a plot of the distribution of the FD of the set of points in a signal having the same Hölder exponent α (which is also called Lipschitz-α, originally introduced in [24] to analyze the homogeneity of multifractal measures, which represent energy dissipation of turbulent fluids) [19]. It was then extended in [25] to multifractal signals. Fig. 1 shows the singularity spectrums for some segmented phonemes from TIMIT database using the FRACLAB tool [26].

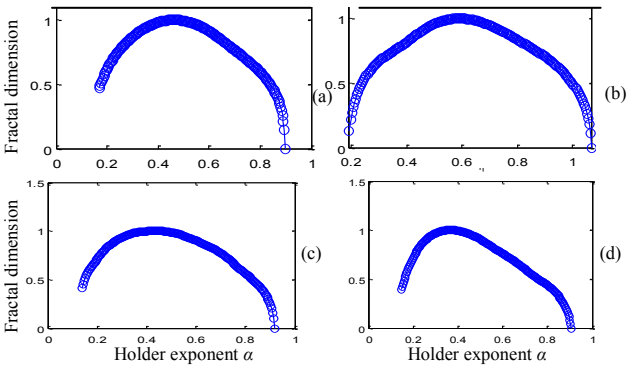

Fig. 1. Singularity spectrum for phonemes (a) /*y*/, (b) /*d*/(plosive), (c) /*ae*/ (vowel) and (d) /*f*/ (fricative).

The plots in Fig. 1 indicate two evidences, *viz.*, 1) within a single speech frame, there exists set of points with *different* Hölder exponents α, which have varying FDs. This means that within the speech frame, one can obtain more than one value of fractal dimension. Hence, use of *multiscale* FD is more appropriate over using *1-D* FD value for the given frame. 2) In addition, the plots for various phonemes, show discriminative property of the singularity spectrum for speech frames. The plots represent the density of the distribution of points with various Hölder exponents. Therefore, it would be beneficial to have a multiscale FD for each frame. As described in [5], the multiscale FD of a *1-D* time series describes variation of the measure of FD over different scales of structuring element. Hence, for a single phoneme, one can have different ranges of FD as will be described in Section 3.1.

Let $f(t)$ be a given speech signal for which disjoint cover is made of its support with intervals of size 'ε'. Then the number of intervals that intersect S_α (*i.e.,* set of all points in $t\in\mathbb{R}$, where the pointwise Lipschitz regularity of $f(t)$ is α),

$$\begin{aligned} &N_\alpha(\varepsilon)\propto\varepsilon,^{-D(\alpha)}\\ &\Rightarrow\log\left(N_\alpha(\alpha)\right)\propto D(\alpha)\log(1/\varepsilon),\\ &\Rightarrow D(\alpha)\propto\frac{\log\left(N_\alpha(\varepsilon)\right)}{\log\left(1/\varepsilon\right)}. \end{aligned} \quad (6)$$

Thus, the singularity spectrum $D(\alpha)$ is FD of S_α and it gives properties of Lipschitz-α singularities that appear at any scale ε.

3.1 FD as feature vectors

The multiscale FD was calculated using the methodology depicted in Section 2. Scales of ε range from *1* to *64* (which correspond to scales of *1/16 ms* to *4 ms*). The value of ε between *11* and *64* are found to be most discriminative. Therefore, *53-D* vectors for each frame are computed. The multiscale FD computed in this way, describe unique properties of the different classes of phonemes. For vowels, values of FD at lower scale (i.e., $\leq$ *1.2 ms*) are low (between *1.3* and *1.6*) owing to less turbulence in vowel production. Furthermore, when scale is increased, FD increases owing to the similarity with the system with increasing signal frequency with constant sampling frequency [5]. In contrast for fricatives (which have high turbulence throughout their production) have a constant high value of FD at *all* scales. For semi-vowels, the value lies at mid-range (*1.5* to *1.7*).

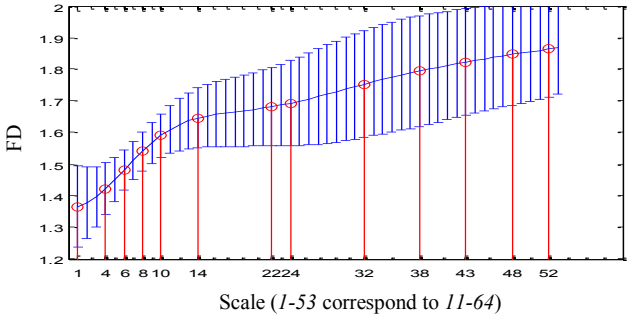

Fig. 2. Plot of phoneme /*iy*/ showing the *13* indices of MFD chosen to be used for recognition used in *FSB* as defined in Sec. 4.2.

This variation in the plots can be exploited as measures similar to the spectral transitivity property of Δ and ΔΔ features while moving from the class of phoneme to the other. In this work, *39-D* feature vector (i.e., *12* MFCC+*1* energy +*13* FD +*13* Δ-FD) is considered, which is compared with MFCC-EDA (where EDA is usual energy-delta-acceleration extensions to MFCC) (*i.e.*, *39-D*) features. The results of the recognition tasks are shown in Table 1. Furthermore, in order to investigate whether the FD, Δ and ΔΔ of MFCCs are equivalent, *48-D* feature vector (i.e., MFCC-EDA + *3*-FD + *3* Δ-FD + *3* ΔΔ-FD) is also considered. Secondly, the choice of specific number of components from *53-D* raw vectors obtained from the morphological operations on the signal has to be dealt with. It should be noted that the additional components should provide better discrimination between different phonemes rather than being redundant. Accordingly, the components are chosen as per the plot shown in Fig. 2, where the coordinates, at which the dynamics for different phonemes were most discriminating, are highlighted.

3.2 Procedure for building frame classification system

The TIMIT database is used to evaluate the performance of features which capture the multiscale nonlinearity from the speech signal in phoneme classification task. TIMIT is provided with a set of manually labeled files (sample-level) for all the utterances. These are converted into frame-level labels as per the specifications used for calculating the various feature vectors. More specifications will be mentioned in Section 4.2. These label files are taken to be *true* labels to be used for training a single hidden layer neural network [27]. The features' discrimination capability is evaluated by testing the model's performance over the unseen cross-validation utterances.

4. EXPERIMENTAL RESULTS

4.1 Data

The *TIMIT* database was used to perform the frame classification experiments [28-29]. The dataset contains *6300* sentences (*10* sentences each from *630* speakers). Out of these, the sentences from '*sa*' category were removed as they were repeated sentences and hence could bias the system. From the remaining utterances, *3696* sentences (provided by *TIMIT*), with *8* utterances spoken by *462* speakers each, were used for training the neural networks and the remaining *full-test* set containing *1344* sentences, with *8* utterances spoken by *168* speakers each (different from those used in training), were used for cross-validating the system's performance.

4.2 Setup

In this work, *4* different MLPs were trained on the *TIMIT* database with following feature vectors,

1. *39-D* feature set A (*FSA*) (i.e., *12* MFCC +*1* Energy + *13* Δ MFCC+ *13* ΔΔ MFCC).
2. *39-D* feature set B (*FSB*) (i.e., *12* MFCC +*1* Energy + *13* FD + *13* ΔFD).
3. *48-D* feature set C (*FSC*) (i.e., *12* MFCC +*1* Energy +*13* Δ MFCC + *13* ΔΔ MFCC + *3* FD +*3* ΔFD +3 ΔΔFD).
4. *48-D* feature set D (*FSD*) (i.e., *15* MFCC +*1* Energy + *16* Δ MFCC + *16* ΔΔ MFCC).

MFCC features are extracted using *25 ms* window length with *10 ms* shift. Similarly, FDs were extracted at frame-level with same specifications. After the FD extraction per frame at various ε values, appropriate coefficients from the *53-D* vectors were augmented with the MFCCs as indicated in Section 3.1 to form FSB and FSC. In particular, for FSC, the 3^{rd}, 6^{th} and 9^{th} coefficients were chosen with their Δs and ΔΔs making up a *48-D* feature vector. *FSD* is constructed as a control feature to ensure that the improvement, if any, obtained with *FSC*, is absolute, and not just due to increase in dimensionality. For *FSD*, in order to obtain a *48-D* MFCC feature vector, the number of filters was increased from *20* to *26* so that while picking the DCT components the *averaging* effect (due to broader filters at higher frequencies) for the higher components is reduced. Single-hidden layer feedforward neural networks (also referred as *multi-layer perceptrons*, or *MLPs* in general) with varying

number of hidden neurons were built as shown in Table 1, to measure the discriminability of the features. The neural network used is taken to be a simple non-linear classifier so that the results obtained are least dependent on the language models (as in HMM-GMM modeling), and is only dependent on the type of features that is presented at its input. For all the features, a context window of *9* frames around the center frame is used. The number of output class is fixed to be *61*, which is the number of phonemes as given in TIMIT for English language. All MLPs are trained using standard back-propagation algorithm, and employ the *newbob* learning rate schedule.

4.2 Results

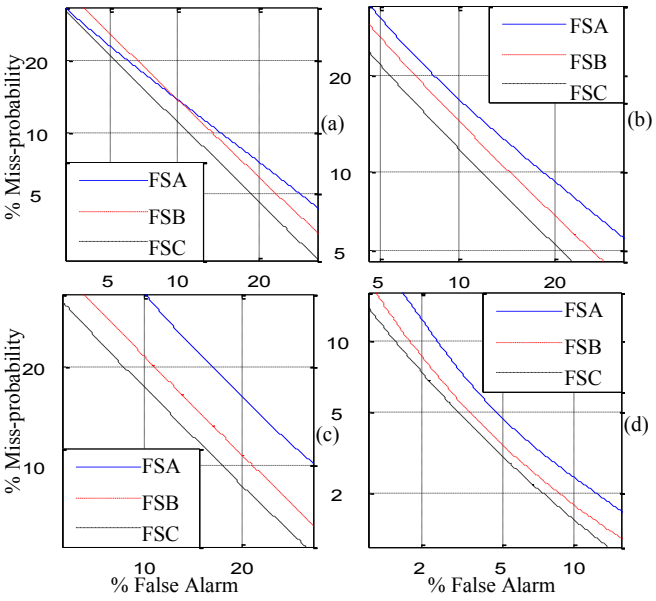

Fig. 3: DET plots for various classifications (a) all *61* classes, (b) *6* classes (c) vowels only and (d) nasals only.

Table 1 shows the training accuracy and CV accuracy of all four features. In addition, the NIST standardized DET [30] are plotted in Fig. 3 for *3* features *FSA*, *FSB* and *FSC* with different groups of output classes. It is evident that *FSC* performed better than *FSA*, *FSB* and *FSD*. For the best performance, i.e., with *1700* neurons, FSC reduces the error rate (*100* % - FCR) by over *1.6* %. FSB are not as discriminative as FSA or FSC, however, they do show promising results. Table 1 shows that *FSD* performs poorer than the *39-D* MFCC (FSA), as increasing the number of filters to obtain higher orders of MFCC, may lead to poorer features which include harmonics. These features may prove ineffective, if not detrimental, for speech recognition task. Hence, we do not include them in the DET curves. Furthermore, from Table 2 at the point of equal error rate (EER), *FSB* is at slight edge over *FSA* which indicates the effectiveness of FD-based features over the Δ and ΔΔ features of MFCC. FD is expected to classify different classes of phonemes efficiently, as shown in Fig. 3(b). Similarly, for the case of vowels, nasals, semi-vowels, plosi-

Feature set		No. of hidden neurons 1000	1500	1700	2048	2300
FSA	Tr. A	71.15	71.36	73.10	73.66	74.03
	CV A	62.14	62.68	62.56	62.30	62.45
FSB	Tr. A	64.49	64.05	64.23	63.91	64.09
	CV A	57.07	57.05	56.99	57.15	57.20
FSC	Tr. A	71.53	73.12	72.41	72.94	73.20
	CV A	**62.26**	61.93	**62.60**	**62.74**	**62.72**
FSD	Tr. A	72.19	72.98	73.38	73.94	74.39
	CV A	61.85	62.35	62.22	62.40	62.55

Table 1: Best frame classification rates (FCRs) obtained for different features with different number of hidden neurons.

	FSA	FSB	FSC
		(Opt. DCF ($\times 10^{-2}$))	
All *61* classes	12.445 (1.37)	12.212 (1.43)	**10.982 (1.39)**
Six classes	13.623 (1.57)	12.651 (1.51)	**11.454 (1.44)**
Vowels	18.125 (1.54)	14.875 (1.38)	**13.33 (1.34)**
Semi-vowels	7.696 (1.24)	7.825 (1.13)	**7.369 (1.02)**
Fricatives	9.044 (1.48)	7.678 (1.26)	**7.652 (1.18)**
Plosives	9.875 (1.38)	10.045 (1.42)	**9.341 (1.38)**
Nasals	4.814 (1.02)	4.179 (0.9)	**3.923 (0.77)**
Silence	3.169 (0.67)	4.535 (0.67)	**3.235 (0.52)**

Table 2: % EER for various classes with different features.

-ves and fricatives, the proposed features perform better than *FSA*. Table 2 indicates an improvement of more than *1.6* % with the proposed features over traditional MFCC feature vectors. Table 2 also shows the minimum Detection Cost Function (DCF) [30] for the three set of features for P_{true} equal to (*1/61*). C_{fa} and C_{miss} were selected to be *1*. The optimum DCFs also show improvement with the proposed features. The number of genuine trials is N_g = *3, 99, 681* (which equal the number of frames in the cross- validation dataset) and number of imposter trials are N_g ×*60*. The curves indicate that the proposed feature outperforms the others at *all* operating points of DET curve.

5. SUMMARY AND CONCLUSIONS

In this paper, effectiveness of FD-based feature is demonstrated for ASR task. Instead of Δ and ΔΔ of the MFCC, FD at various scales and its Δ's were used. The result was an increase in % recognition accuracy as well as significant % EER reduction on the DET curves. This proves that FD captures some additional aspects of the speech signal than the MFCCs, Δ and ΔΔ. To this extent, we have proposed a new feature augmented with MFCC. FD successfully captures the dynamic information from the speech signal at the phoneme-level due to difference in levels of turbulence. Moreover, the transition from one class

of phonemes to the other is effectively detected by the proposed features. However, time complexity required for computing the *64–D multiscale* FD is very high. Our future work would be directed towards finding the robustness of the proposed features under signal degradation conditions.

6. REFERENCES

[1] B. B. Mandelbrot, The Fractal Geometry of Nature, *W. H. Freeman and Co.*, New York, 1983.

[2] V. Pitsikalis and P. Maragos, "Analysis and classification of speech signals by generalized fractal dimension features", in *Speech Communication, vol.* 51, *no.* 12, *pp.* 1206-1223, Elsevier, 2009.

[3] T. F. Quatieri, Dicrete-Time Speech Signal Processing: Principles and Practice, *Pearson Education*, 2004.

[4] D. J. Sinder, "Speech synthesis using an aeroacoustic fricatives model", Ph. D. Thesis, *Rutgers University*, 1999.

[5] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition", in *J. Acoust. Soc. Amer., vol*. 105, *no*. 3, *pp*. 1925-1932, 1998.

[6] P. Henriquez, J. B. Alonso, M. A. Ferrer and C. M. Travieso, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics", in *IEEE Trans. Audio, Speech and Language Process., vol*. 7, *no*. 6, *pp*. 1186-1195, 2009.

[7] P. N. Baljekar, and H. A. Patil, "A comparison of waveform fractal dimension techniques for voice pathology classification" in *Proc of IEEE Int. Conf. Acoustics, Speech and Signal Process.* (ICASSP), Kyoto, Japan, *pp.* 4461-4464, 2012.

[8] F. V. Nelwamondo, U. Mahola, T. Marwola, "Multi-scale fractal dimension for speaker identification systems," *WSEAS Trans. on Systems*, *vol.* 5, *no.* 5, *pp.* 1152–1157, 2006.

[9] F. Martinez, A. Guillamon, J. J. Martinez, "Vowel and consonant characterization using fractal dimension in natural speech," in *Int. Conf. on Nonlinear Speech Process.* (NOLISP), 2003.

[10] L. J. Hadjileontiadis, "A novel technique for denoising explosive lung sounds empirical mode decomposition and fractal dimension filter", in *IEEE Engg. In Medicine and Biology Magazine, vol*. 26, *no*. 1, *pp*. 30-39, 2007.

[11] P. Maragos, "Fractal aspects of speech signals: dimension and interpolation," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Process*. (ICASSP), Toronto, Canada, *pp*. 417–420, 1991.

[12] V. Pitsikalis, I. Kokkinos, P. Maragos, "Nonlinear analysis of speech signals: generalized dimensions and Lyapunov exponents," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003.

[13] V. Pitsikalis and P. Maragos, "Filtered dynamics and fractal dimensions for noisy speech recognition," *IEEE Sig. Process. Lett., vol.* 13, *no.* 11, *pp.* 711–714, 2006.

[14] A. Ezeiza, K. Lopez de Ipina, C. Hernandez, N. Barroso, "Enhancing the feature extraction process for automatic speech recognition with fractal dimensions", in *Cognitive Computation, Springer*, *vol*. 5, *no* 4, *pp*. 545-550, 2013.

[15] M. Zaki, N. J. Shah and H. A. Patil, "Effectiveness of fractal dimension for ASR in low resource language," in *Proc. 9th Int. Symp. Chinese Spoken Lang. Process.* (ISCSLP), Singapore, *pp*. 464-468, 2014.

[16] J. Jimenez, "The contributions of A. N. Kolmogorov to the theory of turbulence," in *Arbor CLXXVIII*, *no*. 704, *pp*. 589-606, 2004.

[17] H. M. Teager and S. M. Teager, " Evidence for nonlinear sound production mechanisms in the vocal tract", in *Speech Production and Speech Modeling*, *NATO Advance Study Institute series D*, *vol*. 55, Bonas, France, *pp*. 241-261, July 1989.

[18] J. F. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view", in *Vocal Fold Physiology, Biomechanics, Acoust., and Phonatory Control*, *pp*. 358-386, 1983.

[19] S. Mallat, A Wavelet Tour of Signal Processing, *Academic Press 2nd ed., Elsevier,* 2006.

[20] U. Frisch, Turbulence: the legacy of A. N. Kolmogorov, *Cambridge University Press*, 1999.

[21] P. Maragos, "Fractal signal analysis using mathematical morphology", Book Chapter in *Advances in Electronics and Electron Physics,* New York, edited by P. Hawkes and B. Kazan, *vol*. 88, *pp.*199-246, 1994.

[22] A. Z. R. Langi, K. Soemintapura, W. Kinsner, "Multifractal processing of speech signals," in *Int. Conf. on Information, Comm. and Signal Process.*, *vol.* 1, *pp.* 527-531, 1997.

[23] D. C. Gonzalez, L. L. Ling, F. Violaro, "Analysis of multifractal nature of speech signals," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, LNCS, *Springer-Verlag,* Berlin Heidelberg, *vol.* 7441, *pp.* 740-748, 2012.

[24] U. Frisch and G. Parisi, Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics, in *Fully Developed Turbulence and Intermittency*, North Holland, 1985.

[25] J. F. Muzy, E. Bacry and A. Arneodo, "The multifractal formalism revisited with wavelets," *Int. J. Bifurcation Chaos*, *vol.* 4, *no.* 2, *pp.* 245-302, 1994.

[26] INRIA (Institute Nationale de Recherche en Informatique et en Automatique) Fraclab: http://fraclab.saclay.inria.fr/ homepage.html. {Last Accessed on Feb. 15th , 2015}.

[27] ICSI Speech Group Tools : QuickNet [Available Online] http :// www. icsi. berkeley. edu/Speech/ icsi-speech-tools. html. {Last Accessed on Feb. 15th , 2015}.

[28] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov Models," in *IEEE Trans. Acoustics, Speech and Signal Process., vol*. 37 *no*. 11, *pp*. 1642-1648, 1989.

[29] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database", *National Institute of Standards and Technology* (NIST), Gaithersburgh, MD, 1988.

[30] A. F. Martin, *et al*., "The DET curve in assessment of detection task performance", in *Proc. EUROSPEECH*, *vol*. 4, *pp*. 1899-1903, 1997.