Effectiveness of Fractal Dimension for ASR in Low Resource Language

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 (MFD) as components of feature vectors for automatic speech recognition (ASR) especially in low resource languages. Speech, which is known to be a nonlinear process, can be efficiently represented by extracting some nonlinear properties such as fractal dimension, from the speech segment. During speech production, vortices (generated due to presence of separated airflow) may travel along the vocal tract and excite vocal tract resonators at the epiglottis, velum, palate, teeth, lips, etc. By Kolmogorov's law, the gradient in energy levels between these vortices produces turbulence. This ruggedness, and in effect, the embedded features of different phoneme classes, can be captured by invariant property of FD. Furthermore, speech is a multifractal, which justifies the use of multiscale fractal dimension as feature components for speech. In this paper, we describe the multifractal nature of speech signal and use this property for automatic phonetic segmentation task. The results show a significant decrease in % EER (≈ 4.2 % from traditional MFCC base features and ≈ 2.5 % from MFCC appended with 1-D fractal dimension). The DET curves clearly show improvement in the performance with the new multiscale fractal dimension-based features for low resource language under consideration.

Index Terms: Automatic phonetic segmentation, multiscale fractal dimension, multifractal, nonlinearities.

1. Introduction
Phonetic segmentation is the task of labeling the given speech waveform into text phonemes accurately which can then be used effectively in various speech technology applications. Labeling the speech data in such a way allows for accurate training of the Hidden Markov Models (HMMs) in Automatic Speech Recognition (ASR) tasks. In addition, labeled speech data is also used to build efficient speech synthesis (both hidden Markov model-based speech synthesis and unit-selection-based speech synthesis) systems. Therefore, correctly labeled data has importance in speech processing. This task of labeling the speech data is mostly done manually for best segmentation results, which is extremely tedious and time-consuming. Therefore, there is a need for labeling or segmenting the speech data automatically without compromising too much on the accuracy.

This automatic segmentation is carried out by the forced Viterbi alignment technique which intends to find the best possible sequence of phonemes given the framewise feature vectors of speech waveform. Naturally, the feature vectors here, which are determined at frame-level, have to be efficient speech representations. Mel Frequency Cepstral Coefficients (MFCCs) are the state-of-the-art standard tool for speech feature extraction process. The MFCCs are very efficient in representing the speech signal in linear settings [1]. However, it does not capture the nonlinear dynamics of speech production mechanism.

The linear source-filter theory of speech production, which assumes that speech is produced only due to the acoustic motion of the air, has been considered to be too stringent in modeling the speech production mechanism [2], [3]. In particular, the model assumes the pressure and velocity field at a given cross-section of the vocal tract to be constant. However, research has shown the presence of jet airflow across the vocal tract which separates the stagnant air in the vocal tract into two different regions with different pressure and velocities. This jet of air stream experiences a viscous force (which is neglected by the linear model), which make the airflow to roll up and create vortices. These vortices move in the opposite direction to that of the jet-stream but with significantly smaller velocity. The vortices are not efficient radiators of speech. However, they do effect the speech production when they come in contact with nonuniform cross-section of the vocal tract. Thus, there exists an aeroacoustic component, which is triggered by the lips, teeth, epiglottis, palate, etc. along with the acoustic component which influences speech production.

Linear models can be used for approximating such nonlinear behavior of speech. However, they require a large data size to encompass such anomalies [5]. This is the main issue with low resourced languages, i.e., scarcity of data. In this study, we use features which are expected to capture the nonlinearity in speech signal so that the amount of data required to efficiently represent speech is feasible. In particular, fractal dimension (FD) has been concatenated with MFCCs to incorporate the nonlinear nature of speech [4], [5].

The use of FD in speech has been a topic of research since as early as studies reported in [6], [7] and has been used in speaker identification tasks, endpoint detection [8], pathological vs. non-pathological classification tasks [9], etc. However, the use of FD in speech segmentation was introduced by [5]. In addition, studies in [4] use 1-D FD measure combined with MFCCs to improve the speech recognition accuracy when applied to low-resourced languages. We argue here that speech being nonstationary, has a multifractal structure and hence one needs to use multiscale FD, as introduced in [5], to account for interscale variations within speech signal and hence finer discrimination between the phoneme classes. For evaluation of the proposed feature vectors, we perform automatic speech recognition (ASR) task and compare the results with standard MFCC feature-based approach. Experimental results are shown on Gujarati (a low resourced Indian language).
2. Fractal Dimension (FD) in Speech

As mentioned earlier, the creation of vortices induce nonlinearities in speech production. These vortices cause the air fluid to form *eddies* generating turbulence in the airflow. According the law to conservation of momentum, we have the Navier-Stokes law for speech production given by [10], [11]:

\[ \rho \left( \frac{\partial u}{\partial t} + u \cdot \nabla u \right) = -\nabla p + \mu \nabla^2 u, \]

where \( \rho \) is the air density, \( p \) is the air pressure, \( u \) is the vector air particle velocity and ‘\( \mu \)’ is the air-viscosity coefficient, assuming negligible flow compressibility. The Reynolds number (Re) is a characterizing parameter for measuring the regularity of aerodynamics. It has been shown in [7] and [12] that speech production consists of dynamics which have a very high value of Re, thereby generating turbulent airflow [5]. In addition, the Kolmogorov law states that velocity field can be modeled as a process \( V(x) \) whose increments has a variance given by [11]:

\[ E \left\{ \left[ V (x+k) - V (x) \right]^2 \right\} = C_2 \frac{2}{\pi} k^{\gamma - 3/2}, \]

where \( C_2 = \frac{2}{\pi} \) is a universal constant, \( k \) 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 [5]. The velocity field which gives rise to turbulence is a power law function. Hence, it does not have a characteristics length scale associated with it in the inertial range of values. Therefore, turbulence would look *self-similar* on magnification at these scales. Furthermore, experiments which indicate the presence of vortices in the airflow indicate quasi-singularity in the statistics of the system, which means that the structure behaves singular until the minimum scale at which smoothness due to viscosity becomes important. Thus, there is a presence of a statistical distribution of singularities over the whole velocity field. For finer analysis of the situation, one needs to separate the different regions of quasi-singularities according to the scales at which they behave singular. Hence, it becomes essential to understand the multiscale characteristics of speech signal which can be efficiently captured by its multiscale fractal dimension [5]. The FD is a robust and efficient parameter for measuring the *roughness* or roughness of any 1-D signal. The next sub-sections give a brief description of the 1-D FD used for comparison and then a brief theory for multiscale FD used for feature extraction process is presented.

2.1. Higuchi’s Fractal Dimension (HFD)

Higuchi’s algorithm along with Katz and modified Range-Scale (R-S) analysis were used in [4], [13] for introducing nonlinear aspects of the data in MFCC feature vectors. However, as indicated in [13], Higuchi’s method for finding the 1-D proves to be very efficient both in computational complexity and accuracy in approximating the actual FD values. Therefore, here we used Higuchi’s algorithm as a benchmark along with the standard MFCC. The FD is calculated for each frame of input speech which is 25 ms with a shift of 10 ms so that it can be directly appended to the end of the 39-D MFCC feature vector (which consists of 12 MFCC + 1 energy + 13-A + 13-D of MFCC). Next, Higuchi’s algorithm is briefly discussed. A more detailed description can be found in [14]. Let the time series of data be given by, \( y(1), y(2), \ldots, y(N) \), which are sampled at regular intervals. Let \( k \) range from 1 to \( k_{max} \), where \( k_{max} \) is taken to be 8. For each value of \( k \), we construct \( k \) new time series in the following way,

\[ y^m_{n} = (y(m), y(m+k), y(m+2k), \ldots, y(m+\text{int}(\frac{N-m}{k})), \]

for \( m=1, 2, \ldots, k \). For each of the time series formed, we calculate the length with appropriate normalization factor, i.e.,

\[ L_{k^m} = \left[ \sum_{j=1}^{\text{int}(\frac{N-m}{k})} \left( y(m+j.k) - y(m+(j-1).k) \right) \right]^\frac{N-1}{\text{int}(\frac{N-m}{k})}. \]

The factor \( \frac{N-1}{\text{int}(\frac{N-m}{k})} \) serves as a normalization factor for different time series with different length. Now, we calculate the mean lengths of the \( k \) time series as,

\[ \left\langle L_{k} \right\rangle = \frac{1}{k} \sum_{m=1}^{k} L_{k^m}, \quad k = 1, 2, 3, \ldots, k_{max}. \]

Now, as in the original box-counting method, we assume here that the lengths obtained in this way have the following relation with the measure of diameter, (here \( k \)),

\[ \left\langle L_{k} \right\rangle \propto k^{-D}. \]

Clearly, if we plot the graph of \( \ln\left( \left\langle L_{k} \right\rangle \right) \) vs \( \ln(k) \), we should get a straight line with a *negative* slope, the absolute value of which being equal to the FD of the time series.

2.2. Multiscale Fractal Dimension (MFD)

FD as defined by Mandelbrot is given by,

\[ D = \lim_{\varepsilon \to 0} \frac{\log(N(\varepsilon))}{\log(1/\varepsilon)}, \]

where \( N(\varepsilon) \) is the number of *compact* planar shapes of size \( \varepsilon \) required to cover the fractal objects under consideration and \( D \) is the FD of the object [15], [17]. However, different speech segments have different levels of turbulence. Hence, ruggedness in the 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 when it comes to implementation [5]. The idea behind the algorithm is to measure the area covered by the object at various scales given by,

\[ A_{\varepsilon}(\varepsilon) = \text{area}(S \oplus eG), \]

where \( S \) is the object under consideration and \( eG \) is the \( e \)-scaled covering element and \( \oplus \) is the nonlinear morphological dilation operation. The fractal dimension, \( D \), is defined as,

\[ D = 2 - \lim_{\varepsilon \to 0} \frac{\log A_{\varepsilon}(\varepsilon)}{\log(1/\varepsilon)}. \]

Eq. (4) and eq. (5) represent continuous case for morphological FD algorithm. The discrete version is given by [16],

\[ A_{\varepsilon}(\varepsilon) = \sum_{n=0}^{N} \left[ (S \oplus \Theta G[n]) \left[ n - (S \Theta G[n]) \right] + O(e^2) \right], \]

\[ S \oplus G[n] = \max \left\{ |S[n+k] + G[k]|, \varepsilon = 1, \right\} \]

\[ S \Theta G[n] = \min \left\{ |S[n+k] - G[k]|, \right\} \]

where \( \Theta \) is the nonlinear morphological erosion operation.
It is to be noted here that the above calculations are for the value of $\varepsilon = 1$. It has been proven [16] that in order to move to the next higher scale of $\varepsilon$, we need to iterate the process again,

$$S \circ G_{\varepsilon+1} = (S \circ G_{\varepsilon}) \circ G, \quad (8)$$
$$S \circ G_{\varepsilon+1} = (S \circ G_{\varepsilon}) \circ G \cdot \quad (9)$$

If we repeatedly perform this for say $\varepsilon = 1, 2, \ldots, N$ times, then we can achieve fractal dimensions with covering elements of size $1/\varepsilon, 2/\varepsilon, \ldots, N/\varepsilon$ seconds.

2.3. Multifractal nature of speech

The multifractal nature of speech signals have been proved in [17] and [18]. The singularity spectrum is a good indicator of the multifractal property of a signal [19]. The singularity spectrum is a plot of the distribution of the FD of the set of points in a signal having the same Holder exponent or Lipschitz regularity $\alpha$. We here show the singularity spectrum for some segmented phonemes such as /aa/ and /s/ in figure 1. The claim is that the speech signal within the frame duration exhibit multifractal nature instead of a monofractal nature. Therefore, it would be more appropriate to use multiscale FD for each frame rather than a 1-D estimate of FD as suggested in [5]. The plots in figure 1 clearly indicate the presence of more than one kind of singularities within each phoneme. As described in [5], the multiscale fractal dimension of a 1-D time series describes the 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 shown in figure 1. The plots show the variation of FD over various values of scales ($\varepsilon$). The multiscale fractal dimension was calculated using the methodology depicted in Section 2.1. Scales of $\varepsilon$ ranging from 1 to 64 (which corresponds to scales of $1/16 \text{ ms}$ to $4 \text{ ms}$) has been used. It has been empirically found that scales from $\varepsilon$ equal to 11 to 64 are found to be most differentiating. Therefore, 53-D row vector for each frame is computed. The plots shown in figure 2 describe unique properties of the different classes of phonemes. For the vowels, values of FD at lower scale $(i.e. \leq 12 \text{ ms})$ are low (between 1.3 to 1.6). This corresponds to the fact that production of vowels has less turbulence due to its inherent periodicity. Furthermore, when scale is increased, FD increases owing to similarity with the system with increasing signal frequency with constant sampling frequency [16]. In contrast, for the case of fricatives (which has high turbulence throughout their production phase) has a constant high value of FD at all scales. For the case of semi-vowels, the value lies at mid-range (between 1.5 to 1.7). This clearly shows the potential of FD to be used in broad classification of phonemes.

3. Experimental Results

3.1. Setup

Three set of features were prepared for measuring the effectiveness of multiscale fractal dimension property of the speech signal. First, simple 39-D MFCC (feature set A (FSA): 12 MFCC + 1 Energy + 13 $\Delta$ + 13 $\Delta\Delta$) as features for forced Viterbi alignment were taken as reference to all the further variations in feature vectors. Second, Higuchi’s 1-D FD (HFD) was calculated for each frame of 25 ms length and a shift of 10 ms. The HFD was observed to be the best among other 1-D FDs in [4]. This 1-D FD is appended to the end of each frame to the corresponding MFCC forming a 40-D feature vector (feature set B (FSB): 12 MFCC + 1 Energy + 13 $\Delta$ + 13 $\Delta\Delta$ + 1 HFD). Next, multiscale FD was calculated as described in Section 2.2 and Section 2.3. It is to be noted here that this 53-D FD alone are insufficient to provide efficient phonetic recognition. Hence, we append selected components from the 53-D FD vector to the 39-D MFCC.

For the purpose of discrimination between different phonetic classes, we select appropriate scale which would provide maximum difference in FD values between different phoneme classes. From the different plots of multiscale FD vs. $\varepsilon$ shown in figure 2, one can notice that for best discrimination between the different phonetic classes, scales corresponding to lower values of $\varepsilon$ need to be chosen. Hence, we choose $\varepsilon = 3,$ 6 and 9 for the purpose. In addition, we take the $\Delta$ and the $\Delta\Delta$ of the 3 FD values and append them to the MFCC. As a result, a 48-D feature vector (feature set C (FSC): 39 MFCC + 3 FD +3 $\Delta$FD + 3 $\Delta\Delta$FD) is created for each frame. It should be noted that, simply an increase in dimension of feature vectors do not lead to improvement in efficiency, as if the additional dimensions were redundant, the system would not be receiving any additional information which would leave the accuracy unchanged, if not rather deteriorated [5]. The performance of system improves only if it captures some additional information which was not with previous features.

3.2. ASR task

For the evaluation of the proposed features, ASR task was performed with the above mentioned features. The process
here is 2-level. In the first step, HTK (HMM-toolkit) is used to determine labels for the speech signal (which are mostly accurate). In the second step, these label files from forced Viterbi alignment of HTK are used to train a neural network. The Quicknet tool was used for neural network training [20]. The neural net at the end of testing step, outputs the a posteriori probabilities (also sometimes called Multi-Layer Perceptrons (MLPs)) to be decoded as recognized labels [21]. In addition, for FSA, 351×1500×52 was taken to be the neural net configuration. For FSB, in a similar way, we take a 360×1300×52 sized neural net. For FSC, we take 432×2000×52 sized neural net. Here, the first value corresponds to the number of inputs (viz., feature vector dimension × number of input frames to take into account the contextual dependence on adjacent frames, i.e., 4 left and 4 right speech frames). The number of hidden neurons is selected to be one which gives maximum efficiency. The final value (i.e., 52 in this case) corresponds to the number of output classes.

3.3. Results

Gujarati speech database is prepared in professional studio environment [22]. The total duration of speech is around 2.5 hrs, with 2 hrs of training and 30 minutes for testing. The results for the ASR task are shown in Table 1 with 95% confidence intervals (CI) indicated [23]. The results show an improvement in recognition accuracy with MFCC + MFD over MFCC and also from 1-D HFD + MFCC. To show the effectiveness of MFD over the other two features in discrimination between wide classes, Table 3 shows the % Equal Error Rates (EER) for the classification of the phonemes first in the 51 total number of classes and then in super-groups of 6 as described in Table 2. The results indicate an improvement of more than 2 % from the 1-D FD-based features and of more than 4 % from traditional MFCC. Table 4 shows the minimum Detection Cost Function (DCF) [24] for the three set of features for two different values of Ptrue, which again show improvement with the proposed features. The NIST standardized DET [24] is used as an additional tool for evaluation of phoneme recognition performance which are plotted in figure 3. The number of genuine trials are 20,842 and number of imposter trials are 10,42,100. The curves indicate that the proposed feature outperforms the others at all operating points of DET curve.

Table 1: Recognition accuracy in % with different features.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Training Accuracy (%)</th>
<th>CV Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSA</td>
<td>77.58</td>
<td>73.28</td>
</tr>
<tr>
<td>FSB</td>
<td>78.34</td>
<td>75.06</td>
</tr>
<tr>
<td>FSC (95% CI)</td>
<td>78.63 (78.39, 78.88)</td>
<td>75.60 (75.01, 76.20)</td>
</tr>
</tbody>
</table>

Table 2: Classification of phonemes in different classes

<table>
<thead>
<tr>
<th>Class</th>
<th>List of Phones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>/a/,/a/,/e/,/i/,/u/,/o/,/ou/,/u/,/ua/</td>
</tr>
<tr>
<td>Semivowels</td>
<td>/i/,/j/,/r/,/r/,/y/</td>
</tr>
<tr>
<td>Fricatives</td>
<td>/s/,/sh/,/h/,/h/,/h/,/s/,/s/,/s/</td>
</tr>
<tr>
<td>Plosives</td>
<td>/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/,/l/</td>
</tr>
<tr>
<td>Nasals</td>
<td>/n/,/ng/,/ma/,/na/,/na/,/na/,/n/</td>
</tr>
<tr>
<td>Silence</td>
<td>/sil/,/sp/</td>
</tr>
</tbody>
</table>

Table 3: % EER for the different features

<table>
<thead>
<tr>
<th>Classes</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All 51 classes</td>
<td>8.34</td>
</tr>
<tr>
<td>FSA</td>
<td>6.71</td>
</tr>
<tr>
<td>FSB</td>
<td>4.12</td>
</tr>
<tr>
<td>FSC</td>
<td>17.41</td>
</tr>
<tr>
<td>Six classes</td>
<td>9.29</td>
</tr>
<tr>
<td>FSA</td>
<td>7.67</td>
</tr>
</tbody>
</table>

Table 4: Minimum Detection Cost Function (DCF)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Minimum Detection Cost Function (DCF)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cfa = Cmiss = 0.5</td>
</tr>
<tr>
<td></td>
<td>Ptrue = 0.5</td>
</tr>
<tr>
<td></td>
<td>Ptrue = 1/51</td>
</tr>
<tr>
<td>FSA</td>
<td>0.0433</td>
</tr>
<tr>
<td>FSB</td>
<td>0.0282</td>
</tr>
<tr>
<td>FSC</td>
<td>0.0195</td>
</tr>
</tbody>
</table>

Figure 3: DET plots for ASR task with three different features (a) for all 51 classes (b) for six broad classes

4. Summary and Conclusions

In this paper, effectiveness of multiscale FD to be used as nonlinear features for ASR task is demonstrated. In particular, the use of nonlinear features in phoneme recognition task for low resourced language is presented. The use of MFD improves the performance of ASR engine with MFCC appended with 1-D HFD, as the speech is known to exhibit multifractal property. The extraction of useful and complementary information from MFD plots led to the building up of a new 48-D feature vector which has enhanced discriminatory properties as compared to the traditional MFCCs and MFCC + HFD features. However, the effectiveness of the proposed features in noisy environment is an important issue which needs to be addressed. Hence, our future work will be directed towards finding the robustness of the proposed features under signal degradation conditions.

5. Acknowledgements

The authors would like to thank Department of Electronics and Information Technology (DeitY), Govt. of India for sponsoring the consortium project, viz., Development of Text-to-Speech Synthesis Systems in Indian Languages. They also thank the consortium leader Prof. Hema A. Murthy (IIT Madras) and the authorities of DA-IICT to support this research work.
6. References


