ON USING FRACTAL FEATURES OF SPEECH SOUNDS IN AUTOMATIC SPEECH RECOGNITION

Petros Maragos † and Alexandros Potamianos *

† Institute for Language & Speech Processing, Margari 22, Athens 11525, GREECE;
and School of E.C.E., Georgia Institute of Technology, Atlanta, GA 30332, USA.
*AT&T Labs–Research, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, U.S.A.

ABSTRACT

The dynamics of air flow during speech production may often result into some small or large degree of turbulence. In this paper, we quantify the geometry of speech turbulence as reflected in the fragmentation of the time signal by using fractal models. We describe an efficient algorithm for estimating the short-time fractal dimension of speech signals based on multiscale morphological filtering and discuss its potential for phonetic classification. We also report experimental results on using the short-time fractal dimension of speech signals at multiple scales as additional features in an automatic speech recognition system using hidden Markov models, which provides a modest improvement in speech recognition performance.

1. INTRODUCTION

The dynamics of speech airflow might create small or large degrees of turbulence during production of speech sounds by the human vocal tract system. Most approaches modeling speech turbulence at the speech waveform level have focused on the random nature of the corresponding signal component. Another important aspect of speech sounds that contain frication or aspiration is the high-degree of geometrical complexity and fragmentation of their time waveforms; due to lack of a better approach, this has been left unmodeled and treated in the past simply as noise. In this paper, we use fractals [1] to model the geometrical complexity of speech waveforms via their fractal dimension, which quantifies the degree of signal fragmentation. First, we provide some motivation and justification from the field of speech aerodynamics for using fractal dimension to quantify the degree of turbulence in speech signals. Further, a simple and efficient algorithm is described for measuring the fractal dimension based on multiscale morphological filtering [3]. Some of our contributions include the measurement and study of the fractal dimension of speech signals in a short-time (phoneme-based) and multiscale framework, which we believe is necessary since speech signals are nonstationary and their fragmentation may vary across different time scales. In this area, we extend the preliminary experiments in [2] by providing measurements averaged over large numbers of phonemic instances from the TIMIT and ISOLET databases. Aa another contribution, we have used the multiscale fractal dimensions of speech segments as additional features in an automatic speech recognition system based on hidden Markov models (HMMs) and found them to offer a modest improvement to the speech recognition performance.

This research was performed while both authors were with the School of ECE, Georgia Institute of Technology, Atlanta, USA. It was partially supported by the US National Science Foundation under Grants MIP-9396301 and MIP-9421677.

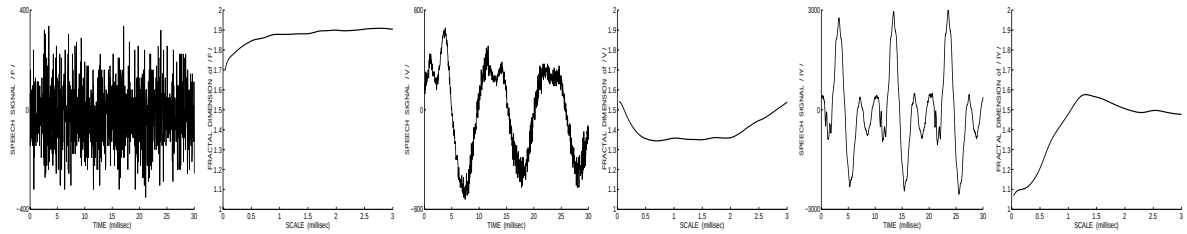
2. SPEECH AERODYNAMICS & FRACTALS

Preservation of momentum in the air flow during speech production yields the Navier-Stokes governing equation [4]

$$\rho(\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \nabla \vec{u}) = -\nabla p + \mu \nabla^2 \vec{u} \qquad (1)$$

where ρ is the air density, p is the air pressure, $\vec{u}$ is the air particle velocity, and μ is the air viscosity coefficient. An important flow parameter is the Reynolds number Re=$\rho UL/\mu$, where U and L are typical velocity and length scales. For the air we have very low μ and hence high Re. This causes the inertia forces (in the left hand side of Eq. (1)) per unit volume to have a much larger order of magnitude than the viscous forces $\mu\nabla^2\vec{u}$. While μ is low and may not play an important role for the speech air flow through the interior of the vocal tract, it is essential for the formation of boundary layers along the tract boundaries and for the creation of vortices. A *vortex* is a region of similar (or constant) vorticity $\vec{\omega}$, where $\vec{\omega} = \nabla \times \vec{u}$. There are several mechanisms for the creation of vortices in the speech air flow: 1) velocity gradients in boundary layers, 2) separation of flow, and 3) curved geometry of tract boundaries. After a vortex has been created, it can propagate downstream and experience twisting, stretching, and diffusion of vorticity [4]. As Re increases (e.g., in fricative sounds or during loud speech), all these phenomena may lead to instabilities and eventually result into *turbulent flow*, which is a 'state of continuous instability' [4] characterized by broad-spectrum rapidly-varying (in space and time) velocity and vorticity. Modern theories that attempt to explain turbulence [4] predict the existence of eddies (vortices with a characteristic size) at multiple scales. According to the energy cascade theory, energy produced by eddies with large size is transferred hierarchically to the small-size eddies which actually dissipate this energy due to viscosity. This multiscale structure of turbulence and several of its geometrical aspects (e.g., shapes of turbulent spots, boundaries of some vortices, shape of particle paths) can be quantified by *fractals* [1].

All the above considerations motivated our use of fractals as a mathematical and computational vehicle to analyze various degrees of turbulence in speech signals. One



(a) (b) (c) (d) (e) (f)

Figure 1: (a,c,e) show speech sounds sampled at 30kHz. (b,d,f) show their corresponding MFDs.

of the main quantitative ideas that we focus on is the fractal dimension of speech signals, because it can quantify their graph's roughness (fragmentation). Since the relationship between turbulence and its fractal geometry or the fractal dimension of the resulting signals is currently not well understood, conceptually equating the amount of turbulence in a speech sound with its fractal dimension may be a simplistic analogy. However, we have found the short-time fractal dimension of speech to be a feature useful for speech sound classification into phonetic classes, segmentation, and recognition.

3. FRACTAL DIMENSIONS OF SPEECH

Let the continuous real-valued function $S(t)$, $0 \leq t \leq T$, represent a short-time speech signal, and let the compact planar set $\mathcal{F} = \{(t, S(t)) \in \mathrm{I\!R}^2 : 0 \leq t \leq T\}$ represent its *graph*. Mandelbrot defines the *fractal dimension* of $\mathcal{F}$ as equal to its Hausdorff dimension D_H; in general, $1 \leq D_H \leq 2$. The signal S is called *fractal* if $D_H > 1$. Another fractal dimension which is closely related to D_H and much easier to compute is the *Minkowski-Bouligand dimension* D. Dilating $\mathcal{F}$ with disks of radius ε and measuring the area $A(\varepsilon)$ of the dilated set, yields D as the constant in the power law $A(\varepsilon) \propto \varepsilon^{2-D}$ as $\varepsilon \to 0$, which $A(\varepsilon)$ obeys if $\mathcal{F}$ is fractal. D is identical to the box counting dimension in continuous time but is more robust to compute in discrete time [3]; henceforth, we shall use D as the 'fractal dimension'.

In continuous time, D will not change if we replace the disks in covering $\mathcal{F}$ with other compact planar shapes B [3]. Thus, if $\varepsilon B = \{\varepsilon b : b \in B\}$ and $A_B(\varepsilon) = \mathrm{area}(\mathcal{F} \oplus \varepsilon B)$ where $\oplus$ is the morphological set dilation,

$$D = \lim_{\varepsilon \to 0} \frac{\log[A_B(\varepsilon)/\varepsilon^2]}{\log(1/\varepsilon)}. \qquad (2)$$

For reducing the computational complexity, it is desirable to obtain the area $A_B(\varepsilon)$ by using 1D operations on $S(t)$. Thus, if the function $G_\varepsilon(t) = \sup\{y \in \mathrm{I\!R} : (t, y) \in \varepsilon B\}$ is the top boundary of εB and if $S \oplus G_\varepsilon$ and $S \ominus G_\varepsilon$ are the morphological function dilation and erosion of S by G at scale ε, then [3]

$$A_B(\varepsilon) = \int_0^T [S \oplus G_\varepsilon(t) - S \ominus G_\varepsilon(t)]dt + O(\varepsilon^2). \qquad (3)$$

These signal dilations and erosions create an area-strip as a layer either covering or being peeled off from the graph of the speech signal at various scales.

For a discrete-time finite-length speech signal $S[n]$, $n = 0, 1, \ldots, N$, we use covers at discrete scales $\varepsilon = 1, 2, \ldots$, and restrict the function $G[n]$ (at scale $\varepsilon = 1$) to have a centered 3-sample support and only two possible shapes: a *triangle*, defined by $G_t[-1] = G_t[1] = 0$ and $G_t[0] = h \geq 0$, or a *rectangle*, defined by $G_r[-1] = G_r[0] = G_r[1] = h \geq 0$. This yields the following scale-recursive algorithm [3]:

$$\begin{aligned} S \oplus G[n] &= \max_{-1 \leq k \leq 1} \{S[n+k] + G[k]\}, & \varepsilon = 1 \\ S \oplus G_{\varepsilon+1} &= (S \oplus G_\varepsilon) \oplus G, & \varepsilon \geq 2. \end{aligned} \qquad (4)$$

where $\varepsilon = 1, 2, 3, \ldots, \varepsilon_{max}$. Likewise for the multiscale erosions $S \ominus G_\varepsilon$. Next, we compute the areas $A_B[\varepsilon]$ by replacing the $\int_0^T$ in (3) with summation $\sum_{n=0}^{N}$. Finally, we fit a straight line using least-squares to the plot of $(\log A_B[\varepsilon]/\varepsilon^2, \log 1/\varepsilon)$. The slope of this line gives us the fractal dimension of S. As height $h = G[0]$, we set $h = 0$, which makes the algorithm faster and invariant to affine transformations in the signal's range.

For real-world signals with some fractal structure, the assumption of a constant D at all scales ε may not be true. Hence, instead of a global dimension, we estimate the *multiscale fractal dimension* MFD[ε], which for each ε is equal to the slope of a line segment fitted via least-squares to the log-log plot over a moving window $\{\varepsilon, \varepsilon + 1, \ldots, \varepsilon + 9\}$ of 10 scales.

Fig. 1 shows 30 ms segments of unvoiced fricative, voiced fricative, and vowel speech sounds extracted from words spoken by a male speaker and sampled at 30 kHz ($N = 900$) together with their corresponding profiles of MFD[ε] for scales $\varepsilon = 1, .., 90$. We have conducted many experiments similar to the ones shown in Fig. 1, from which we conclude the following: 1) Unvoiced fricatives (/f/, /th/, /s/), affricates, stops (during their turbulent phase), and some voiced fricatives like /z/ have a high fractal dimension $\in [1.6, 1.9]$ at all time scales (mostly constant at scales > 1 ms), consistent with the turbulence phenomena present during their production. 2) Vowels at small scales (< 0.1 ms) have a small fractal dimension $\in [1, 1.3]$. This is consistent with the absence or small degree of turbulence (e.g., for loud or breathy speech) during their production. However, at scales $> 2 - 3$ ms, i.e., at scales of the same order as the distance between the major consecutive peaks in the speech waveform their fractal dimension increases appreciably. 3) Some voiced fricatives like /v/ and /th/ have a mixed behavior. If they do not contain a fully developed turbulence state their fractal dimension is medium-to-high $[1.3, 1.6]$ at scales < 0.1 ms, increases at large scales > 3 ms (for the same reasons as for vowels), and may decrease for intermediate scales. Overall, their dimension is high (> 1.6), although often somewhat lower than the dimension of their

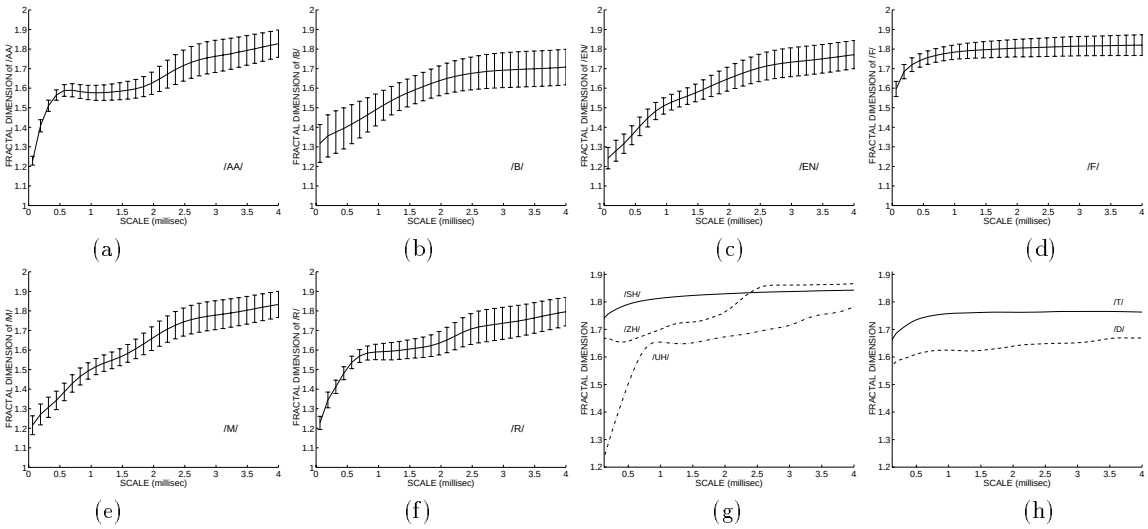


Figure 2: (a,b,c,d,e,f): Mean and standard deviation (error bars) of the multiscale fractal dimension distribution for the phonemes /aa/, /b/, /en/, /f/, /m/, /r/ calculated from the TIMIT database. Also comparison of mean MFD for (g) phonemes /sh/, /zh/, /uh/ and (h) phonemes /t/, /d/. (20 ms analysis window, updated every 10 ms).

unvoiced counterparts. Thus, for normal conversational speech, we have found that the short-time fractal dimension D (computed over $\sim 10-30$ ms frames and evaluated at a scale < 0.1 ms) can roughly distinguish three classes of speech sounds: (i) vowels (small D), (ii) low-turbulence voiced fricatives, e.g., /v/,/th/ (medium D), and (iii) unvoiced fricatives, high-turbulence voiced fricatives, stops, and affricates (large D). However, for loud speech (where the air velocity and Re increase, and hence turbulence occurs more often) or for breathy voice (especially for female speakers) the fractal dimension of several speech sounds, e.g. vowels, may increase. In general, the fractal dimension estimates may be affected by several factors including a) the time scale, b) the specific discrete algorithm, and c) the speaking style. Therefore, we do not assign any particular importance to the absolute estimates but only to their average ranges for classes of speech sounds and to their relative differences.

The short-time MFD computed over scales from 1/16 to 4 ms and averaged over multiple phonemic instances is shown in Fig. 2. For each phoneme the mean and standard deviation of the MFD is computed from 200 instances (100 from male and 100 from female speakers) of each phoneme in the TIMIT database. These multiple averaging experiments verify our previous claims that, for normal conversational speech, the short-time fractal dimension D in small scales can help discriminate among broad phonemic classes. Note that the standard deviation of the MFD distribution is typically smaller for D computed over smaller time scales (< 1 ms), with the exception of the phoneme /b/. Further, the differences among the average fractal dimensions are larger for smaller (< 1 ms) time scales. Figure 2(g) compares the average MFD for the unvoiced fricative /sh/, the corresponding voiced fricative /zh/ and the vowel /uh/. Clearly the small and medium-scale fractal dimension measurement is smaller for voiced than for unvoiced sounds. Further, MFD is able to discriminate between voiced and unvoiced plosives produced with identical vocal tract configuration (thus having very similar short–time spectral envelopes), i.e., /p/ and /b/, /t/ and /d/ e.t.c. For example, Fig. 2(h) shows the average MFD for the voiced–unvoiced plosive pair /d/ and /t/. Again the MFD is smaller for the voiced /d/ than for the unvoiced /t/. The discriminative power of the fractal dimension for fricatives and plosives, where traditional spectral features are inadequate, could be a valuable asset for speech recognition as discussed next.

4. AUTOMATIC SPEECH RECOGNITION

Here we attempt to incorporate the fractal dimension in a hidden Markov model (HMM)–based speech recognizer; mixtures of Gaussian distributions are used to model the observation probabilities for each HMM state.

To successfully incorporate a feature in a pattern classifier the new features must contain if possible only information *relevant* to the discrimination task, i.e., not be redundant or irrelevant. The fractal dimension of a speech signal is defined in this paper to be a 2D distribution in time and scale. The main issue is how to represent this 2D distribution so that it fits in the HMM framework. The feature vectors used in speech recognition are typically computed over a 20-30 ms window and are updated every 5-10 ms. Fractal dimension is a feature with high temporal resolution thus it might be advantageous to avoid over-smoothing. An 8 ms averaging window (updated every 10 ms) was used to compute the fractal features in this paper. The 'standard' speech recognition features (i.e., cepstrum and energy) were computed using a 20 ms window.

The second issue to be resolved is the dimensionality of the fractal feature vector. Smaller dimensionality presents a computational advantage but comes with a performance tradeoff if relevant information is lost during the dimensionality reduction process. It is clear from Fig. 2 that the fractal dimensions of adjacent scales are highly correlated. Further, the fractal dimension of large scales (> 1.5 ms) provide little information relevant to the discrimination task at hand. Various empirical procedures exist for decorrelating a feature vector. We chose the sim-

Table 1: Word Percent Correct for the E-set Recognition Task using 5–Mixture Gaussians per HMM State.

$\{E, C_1..C_{12}, \Delta E, \Delta C_1..\Delta C_{12}\}$	$\{E, C_1..C_{12}, \Delta E, \Delta C_1..\Delta C_{12}\}$ + $\{D_1, \Delta D_1\}$	$\{E, C_1..C_{12}, \Delta E, \Delta C_1..\Delta C_{12}\}$ + $\{D_1, D_{11}, \Delta D_1, \Delta D_6, \Delta D_{11}, \Delta D_{16}\}$
81.2%	83.5 %	84.5%

Table 2: Word Percent Correct for the E-set Recognition Task.

Features / Models	$\{E, C, \Delta E, \Delta C, \Delta\Delta E, \Delta\Delta C\}$	$\{E, C, \Delta E, \Delta C, \Delta\Delta E, \Delta\Delta C\}$ + $\{D, \Delta D\}$
5–mixture Gaussians	85.6 %	86.3%
10–mixture Gaussians	88.6 %	88.9%

plistic approach of sparsely sampling the low-end of fractal scales (< 1 ms).

The feature vector augmented with fractal features as described above was applied to the speech recognition task of the highly confusable e-set consisting of the following spoken letters: b, c, d, g, p, t, v, z. The e-subset of the ISOLET database consists of 2700 word occurrences sampled at 16 kHz [5]. The HMM-based HTK recognition package was used for all experiments [6]. A hold-one-out ("round-robin") procedure was used during training so that all 2700 words were available for testing.

The 'standard' feature set consisted of the energy E, the first twelve cepstrum coefficients $C_1..C_{12}$ computed from a mel filterbank [7] and their first time derivatives ΔE and $\Delta C_1..\Delta C_{12}$. The 'standard' feature vector was augmented by the fractal dimension at scale one $D_1 =$ MFD[1] and its first time derivative ΔD_1. Scale one corresponds to a time scale of 1/16 ms. The fractal features are assumed to be independent of the 'standard' features and to belong in separate probability 'streams'. Five-state left-right hidden Markov models were used in these experiments. As shown in Table 1, combining the 'standard' and the fractal features gives a modest 12% reduction in the word error rate over using the 'standard' features alone. Further improvement is achieved when the higher-scale fractal dimensions (scales 6, 11 and 16, corresponding to time scales of 0.38, 0.69 and 1 ms) are used in addition to D_1 as shown in the third column of Table 1; this yields an error reduction of 18%. Further augmentation of the fractal feature vector has not shown experimentally any performance improvement. Henceforth, we refer to the feature vector consisting of $\{D_1, D_{11}, \Delta D_1, \Delta D_6, \Delta D_{11}, \Delta D_{16}\}$ as the 'fractal' feature vector.

Next, we attempted to improve overall performance by augmentation of our feature set with the second time derivatives of the energy and cepstrum features $\{\Delta\Delta E, \Delta\Delta C_1..\Delta\Delta C_{12}\}$ and by doubling the complexity of the HMM models, i.e., using 10 instead of 5 Gaussian distributions per mixture per state. As shown in Table 2, as the complexity of the models and/or the dimensionality of the 'standard' features increases the improvement in performance achieved by using the fractal features becomes marginal.

Preliminary experiments on general phoneme recognition tasks have shown similar performance improvements when the 'standard' feature vector was augmented with fractal features. Overall, we have found that, fractal features can provide modest improvement to recognition performance with a small increase in the dimensionality of the feature vector.

5. REFERENCES

[1] B. B. Mandelbrot, *The Fractal Geometry of Nature*, NY: W.H. Freeman, 1982.

[2] P. Maragos, "Fractal Aspects of Speech Signals: Dimension and Interpolation", *Proc. ICASSP-91*, Toronto, Canada, May 1991, pp. 417–420.

[3] P. Maragos, "Fractal Signal Analysis Using Mathematical Morphology", in *Advances in Electronics and Electron Physics*, vol. 88, P. Hawkes and B. Kazan, Eds., Academic Press, 1994, pp. 199–246.

[4] D. J. Tritton, *Physical Fluid Dynamics*, Oxford Univ. Press, 1988.

[5] R. Cole, Y. Muthusamy, and M. Fanty, "The ISOLET spoken letter database," Tech. Rep. CSE 90-004, Oregon Grad. Inst. Science & Technology, Mar. 1990.

[6] S. Young et al, *The HTK Book*. Cambridge Research Lab: Entropics, 1995.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 357–366, 1992.