FRACTAL DIMENSIONS OF SPEECH SOUNDS: COMPUTATION AND APPLICATION TO AUTOMATIC SPEECH RECOGNITION *

Petros Maragos[1] and Alexandros Potamianos[2]

[1] Department of Electrical & Computer Engineering,

National Technical University of Athens,

Zografou 15773, Athens, Greece.

Email: maragos@cs.ntua.gr

[2] AT&T Labs, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, U.S.A.

Email: potam@research.att.com

Accepted for publication in the *Journal of the Acoustical Society of America*

First Submission: September 19, 1996. Revised: March 1998. Accepted: September 1998.

Running title: *Fractal Dimensions of Speech Sounds*

Received: _____

*The first part of this work was performed while both authors were at the School of E.C.E., Georgia Institute of Technology, Atlanta, GA 30332, U.S.A., and was supported by the US National Science Foundation under Grants MIP-9396301 and MIP-9421677. The second part was done while P. Maragos was at the Institute for Language & Speech Processing, Athens, Greece.

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 speech segmentation and 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 provide a modest improvement in speech recognition performance.

PACS numbers: 43.72.Ar, 43.72.Ne

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. Static airflow and acoustic characteristics of turbulent speech, e.g., fricative and stop sounds or sounds with aspiration, have been studied by several researchers; references and related discussion can be found in Fant (1970), Flanagan (1972) and Stevens (1971). While the majority of work in this area has associated turbulence in speech mainly with consonants, it is also possible to have vowels uttered with some (speaker-dependent) amount of aspiration which adds to them some small degree of turbulence. To produce natural-sounding synthetic speech it was judged necessary to simulate an aspiration noise source in speech synthesis systems (Klatt, 1987).

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 the theory of fractals (Mandelbrot, 1982) to model the geometrical complexity of speech waveforms via their fractal dimension, which quantifies the degree of signal fragmentation. In Section I, 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. In Section II, a simple and efficient algorithm is described for measuring the fractal dimension. The algorithm is based on multiscale nonlinear operators of morphological filtering

that iteratively expand and contract the signal's graph (Maragos, 1994; Serra, 1982). 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 Maragos (1991) by providing measurements averaged over large numbers of phonemic instances from the TIMIT and ISOLET databases. Another contribution is to the field of speech recognition: the multiscale fractal dimensions of short-time speech segments are used as additional features in an automatic speech recognition system based on hidden Markov models (HMMs). As discussed in Section III, the fractal features can offer a modest improvement to the performance of HMM-based speech recognizers.

I Speech Aerodynamics and Fractals

Conservation of momentum in the air flow during speech production yields the Navier-Stokes governing equation (Tritton, 1988)

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

where ρ is the air density, p is the air pressure, $\vec{u}$ is the (vector) air particle velocity, and μ is the (assumed constant) air viscosity coefficient. It is assumed that flow compressibility is negligible [valid since in speech flow (Mach numbers)$^2 \ll 1$] and hence $\nabla \cdot \vec{u} = 0$. An important parameter characterizing the type of flow is the Reynolds number Re=$\rho UL/\mu$, where U is a velocity scale for $\vec{u}$ and L is a typical length scale, e.g., the tract diameter. 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 $\vec{\omega} = \nabla \times \vec{u}$. Vortices in the air flow have been experimentally found above the glottis by Teager and Teager (1990) and Thomas (1986) and theoretically predicted by Kaiser (1983), Teager and Teager (1990), and McGowan (1988) using simple geometries. There are several mechanisms for the creation of vortices: 1) velocity gradients in boundary layers, 2) separation of flow, which can easily happen at cavity inlets due to adverse pressure gradients (see Kaiser (1983), and Teager and Teager (1990) for experimental evidence for separated flow during speech production), and 3) curved geometry of tract boundaries, where due

to the dominant inertia forces the flow follows the curvature and develops rotational components. After a vortex has been created, it can propagate downstream as governed by the vorticity equation (Tritton, 1988)

$$\frac{\partial \vec{\omega}}{\partial t} + \vec{u} \cdot \nabla \vec{\omega} = \vec{\omega} \cdot \nabla \vec{u} + \nu \nabla^2 \vec{\omega} \quad , \quad \nu = \mu / \rho \tag{2}$$

The term $\vec{\omega} \cdot \nabla \vec{u}$ causes vortex twisting and stretching, whereas $\nu \nabla^2 \vec{\omega}$ produces diffusion of vorticity. As Re increases (e.g., in fricative sounds or during loud speech), all these phenomena may lead to instabilities and eventually result in *turbulent flow*, which is a 'state of continuous instability' (Tritton, 1988) characterized by broad-spectrum rapidly-varying (in space and time) velocity and vorticity. The transition to turbulence during speech production may occur for lower Re closer to the glottis because there is an air jet flowing out from the vocal cords and for jets turbulence starts at much lower Re than for flows attached to walls (as is the case downstream in the vocal tract).

Modern theories that attempt to explain turbulence (Tritton, 1988) predict the existence of eddies (vortices with a characteristic size λ) at multiple scales. According to the energy cascade theory, energy produced by eddies with large size λ (of the order of the boundary layer thickness) is transferred hierarchically to the small-size eddies which actually dissipate this energy due to viscosity. A related result is the Kolmogorov law

$$E(k, r) \propto r^{2/3} k^{-5/3} \tag{3}$$

where $k = 2\pi/\lambda$ is the wavenumber in a finite nonzero range, r is the energy dissipation rate, and $E(k, r)$ is the velocity wavenumber spectrum, i.e., Fourier transform of spatial correlations. This multiscale structure of turbulence can in some cases be quantified by *fractals*. Mandelbrot (1982) and others have conjectured that several geometrical aspects of turbulence (e.g., shapes of turbulent spots, boundaries of some vortex types found in turbulent flows, shape of particle paths) are fractal in nature. We may also attempt to understand aspects of turbulence as cases of chaos. Specifically, chaotic dynamical systems converge to attractors whose sets in phase space or related time-series signals can be modeled by fractals; references can be found in the survey by Peitgen et al. (1992). Now there are several mechanisms in high-Re speech flows that can be viewed as routes to chaos; e.g., vortices twist, stretch, and fold (due to the bounded tract geometry) (Tritton, 1988; Mandelbrot, 1982). This process of twisting, stretching, and folding has been to found in low-order nonlinear dynamical systems to give rise to chaos and fractal attractors.

All the above theoretical considerations and experimental evidence and the fact that the speech signal is produced by a nonlinear dynamical system that often generates small or large degrees of

turbulence motivated our study of its fractal aspects. In this paper, we use fractals as a mathematical and computational vehicle to analyze various degrees of turbulence in speech signals. One 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). Because the relationship between turbulence and its fractal geometry or the fractal dimension of the resulting signals is currently not well understood, in this paper, we conceptually equate the amount of turbulence in a speech sound with its fractal dimension. Although this may be a somewhat simplistic analogy, 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.

II Fractal Dimensions of Speech

A Preliminaries on Fractal Dimensions

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

$$\mathcal{F} = \{(t, S(t)) \in \mathbb{R}^2 : 0 \leq t \leq T\} \tag{4}$$

represent its *graph*. Mandelbrot (1982) 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 its graph is a fractal set, i.e., if D_H strictly exceeds 1 (= the topological dimension of $\mathcal{F}$). Next we discuss two other dimensions closely related to D_H.

Minkowski-Bouligand dimension D_M: This is based conceptually on Minkowski's idea of finding the length of (possibly irregular) curves $\mathcal{F}$: Dilate $\mathcal{F}$ with disks of radius ε by forming the union of these disks centered at all points of $\mathcal{F}$ and thus create a 'Minkowski cover'. Find the area $A(\varepsilon)$ of the dilated set, and set its multiscale length equal to $\lim_{\varepsilon \to 0} L(\varepsilon)$, where $L(\varepsilon) = A(\varepsilon)/2\varepsilon$. Then D_M is the constant D in the power law $L(\varepsilon) \propto \varepsilon^{1-D}$ as $\varepsilon \to 0$, which $L(\varepsilon)$ obeys if $\mathcal{F}$ is fractal.

Box Counting Dimension D_B: Partition the plane with a grid of squares of side ε and count the number $N(\varepsilon)$ of squares that intersect the curve. Then the box dimension is obtained by replacing the Minkowski cover area with the box cover area, i.e., it is equal to $D_B = \lim_{\varepsilon \to 0} \log[N(\varepsilon)]/\log(1/\varepsilon)$.

In general, $1 \leq D_H \leq D_M = D_B \leq 2$. In this work we focus only on D_M which we shall henceforth call the 'fractal dimension' D because: i) It is closely related to D_H and hence able to quantify the fractal aspects of a signal. ii) It coincides with D_H in many cases of practical interest. iii) It is much easier to compute than D_H. iv) It will be applied to sampled signals where most

approaches can yield only approximate results. v) D_M can be more robustly estimated than D_B, which suffers from uncertainties due to the grid translation or its spacing ε relative to the signal's amplitude. Note that, $D_B = D_M$ in the continuous-time case, but they correspond to two different algorithms (with different performances) for sampled signals.

B Morphological Covering Algorithm

As shown by Maragos and Sun (1993), and Maragos (1994), D will not change if we replace the disks in the Minkowski cover of $\mathcal{F}$ with other compact planar shapes B. Thus, if $\varepsilon B = \{\varepsilon b : b \in B\}$ is an ε-scaled shape B, we can obtain multiscale multishape area distributions

$$A_B(\varepsilon) = \text{area}(\mathcal{F} \oplus \varepsilon B) \tag{5}$$

where $\mathcal{F} \oplus \varepsilon B$ is the set resulting from the morphological set dilation of $\mathcal{F}$ by εB:

$$\mathcal{F} \oplus \varepsilon B = \{z + \varepsilon b \in \mathbb{R}^2 : z \in \mathcal{F} \ , \ b \in B\} \tag{6}$$

The infinitesimal order of the multiscale area function yields the fractal dimension of $\mathcal{F}$, i.e.,

$$D = 2 \ - \ \lim_{\varepsilon \to 0} \frac{\log[A_B(\varepsilon)]}{\log(\varepsilon)}. \tag{7}$$

Assuming now that $A_B(\varepsilon) \propto \varepsilon^{2-D}$ as $\varepsilon \to 0$ yields that

$$\log[A_B(\varepsilon)] = (2 - D) \log(\varepsilon) \ + \ \text{constant} \ \ , \ \ \text{as} \ \ \varepsilon \to 0. \tag{8}$$

Thus, in practice D can be estimated by least-squares fitting a straight line to and measuring the slope of the plot of the data $\log[A_B(\varepsilon)]$ versus $\log(\varepsilon)$.

Implementing the set dilation $\mathcal{F} \oplus \varepsilon B$ involves representing the signal graph $\mathcal{F}$ as a binary image signal and dilating this binary image. This, however, two-dimensional processing of a one-dimensional signal $S(t)$, on the one hand is unnecessary and on the other hand increases the requirements in storage space and the time complexity for implementing the covering method. Thus, for purposes of computational efficiency, it is desirable to obtain the area $A_B(\varepsilon)$ by using *one-dimensional operations* on $S(t)$. Toward this goal, let us consider first the morphological dilation $\oplus$ and erosion $\ominus$ of the signal $S(t)$ by a real-valued function $G(t)$ defined as

$$\begin{aligned} (S \oplus G)(t) &= \sup_x \{S(x) + G(t - x)\} \\ (S \ominus G)(t) &= \inf_x \{S(x) - G(x - t)\} \end{aligned}$$

These signal dilation and erosion are one-dimensional nonlinear signal operators whose computational structure is similar to a convolution and correlation, respectively. Then (Maragos and Sun, 1993; Maragos, 1994), if we select as B any compact single-connected and symmetric planar set and if we define

$$G_\varepsilon(t) = \sup\{y \in \mathbb{R} : (t, y) \in \varepsilon B\} \tag{9}$$

as the function (structuring element) whose graph is the top boundary of εB, we obtain

$$A_B(\varepsilon) = \int_0^T [(S \oplus G_\varepsilon)(t) - (S \ominus G_\varepsilon)(t)]dt + O(\varepsilon^2). \tag{10}$$

where $S \oplus G_\varepsilon$ and $S \ominus G_\varepsilon$ are the dilation and erosion of S by G at scale ε. Thus, instead of creating the cover of a one-dimensional signal by dilating its graph in the plane by a set B (which means two-dimensional processing), $S(t)$ can be transformed via an erosion and a dilation by functions G_ε. These 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. We refer to this whole approach as the *morphological covering method.*

Discrete Covering method: To adapt our previous discussion to the case of 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 B becomes a finite set of pixels in the discrete plane. If we restrict the discrete set B to be convex and of radius=1 then the corresponding function $G[n]$ (at scale $\varepsilon = 1$) is restricted 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 $G_r[-1] = G_r[0] = G_r[1] = h \geq 0$. The height h is allowed to vary and match the amplitude range of the signal S. The main result in the discrete case is the following scale-recursive algorithm (Maragos, 1994):

$$\begin{aligned}
S \oplus G[n] &= \max_{-1 \leq k \leq 1} \{S[n+k] + G[k]\}\,, \quad \varepsilon = 1 \\
S \ominus G[n] &= \min_{-1 \leq k \leq 1} \{S[n+k] - G[k]\}\,, \\
S \oplus G_{\varepsilon+1} &= (S \oplus G_\varepsilon) \oplus G\,, \qquad\qquad \varepsilon \geq 2 \\
S \ominus G_{\varepsilon+1} &= (S \ominus G_\varepsilon) \ominus G\,,
\end{aligned} \tag{11}$$

where $\varepsilon = 1, 2, 3, \ldots, \varepsilon_{max}$. For $n = 0, N$ the local max/min operations take place only over the available samples. Next, we compute the areas $A_B[\varepsilon]$ by replacing the $\int_0^T$ in (10) with summation $\sum_{n=0}^N$. Finally, we fit a straight line using least-squares to the plot of $(\log A_B[\varepsilon], \log \varepsilon)$. The slope of this line is an estimate of $2 - D$ and gives us the fractal dimension of S. 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 of (8) over a short moving window $\{\varepsilon, \varepsilon+1, \ldots, \varepsilon+w\}$ of w scales, where in practice w ranges from 3 to 10. Throughout this paper we have used $w = 10$.

The height $h = G[0]$ of the structuring function G is important because it controls the finesse or coarseness of the multiscale area measurements. A good practical rule is to set h less than or equal to the signal's dynamic range divided by the number of its samples. We experimentally observed that this rule performs very similarly to the case $h = 0$. When $h = 0$, i.e., when the function G becomes a flat function corresponding to B being a horizontal segment, Dubuc et al. (1989) have shown that the fractal dimension can still be computed from the above covering method (for continuous-time signals). This case has two advantages: the erosions/dilations can be performed faster, and the algorithm yields local fractal dimensions that are invariant to any affine transformation $S(t) \mapsto aS(t) + b$ of the amplitude range ($a > 0$). Therefore, we henceforth select $h = 0$.

Applying the morphological covering method on a *sampled* signal incurs a discretization error in the estimated fractal dimension. Namely, the graph of the sampled signal has lost some of the degree of fragmentation inherent in the graph of the continuous-time signal or presents a distorted view of the geometry of the continuous-time graph due to a small number of available samples and/or a small number of available scales. Thus, the error depends on the sampling frequency and the essential bandwidth (i.e., the frequency range containing most of the area under the square amplitude spectrum of the signal). Analytic formulas for the error currently do not exist, but experimenting with special cases could be instructive. We have applied the above discrete flat covering algorithm to the particular case of a sampled sine and measured the error between the estimated fractal dimension (at scale $\varepsilon = 1$) versus the theoretically correct value of 1. By analyzing sampled sine signals over a fixed finite time interval we experimentally observed that, for a fixed sampling frequency, increasing the frequency of the sine increases the dimension error. Further, for a fixed frequency of the sine, increasing the sampling frequency decreases the error. We conjecture that this error depends on the oversampling ratio, i.e., the ratio of the sampling frequency divided by the sine frequency. As indicative values from this (experimentally found) error law, with $w = 5$, to obtain an estimated dimension ≤ 1.1 (i.e., an error $\leq 10\%$) we need an oversampling ratio of at least 20:1, whereas an estimated dimension ≤ 1.01 (i.e., an error $\leq 1\%$) requires an oversampling ratio of at least 70:1. At present we have no such guidelines for more general signals, but increasing

the sampling frequency decreases the error.

C Experiments on Measuring Fractal Dimension of Speech Signals

Fig. 1 shows 30 ms segments of an unvoiced fricative, a voiced fricative, and a vowel speech sound 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 $\varepsilon = 1, .., 120$. This range of ε corresponds to time scales from 1/30 ms to 4 ms. The reason for using a higher than usual sampling rate is to approximate as closely as possible the geometrical roughness of the continuous-time speech signal and decrease the discretization error in the estimated fractal dimension. However, similar results have been observed at lower sampling rates (10 and 16 kHz) and are reported in the following sections. 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. Here we observe a phenomenon similar to the previously mentioned increase of the estimated fractal dimension of a fixed time segment of a sine when the sampling frequency remains constant and the sine frequency increases. 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 > 5 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 unvoiced counterparts. Thus, 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). Thus, the fractal dimension consistently quantifies the well-known fact that the geometrical fragmentation of the waveforms of these three speech classes increases in the same order. 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 significantly increase. In general, the fractal dimension estimates may be affected by several factors including a) the time scale, b) the specific discrete algorithm (usually most algorithms for sampled signals underestimate the true fractal dimension since some signal's roughness has been lost during sampling), 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.

We also used D estimated at a single small time scale, i.e., MFD$[\varepsilon = 1]$, as a short-time feature for purposes of speech segmentation and for signaling important events along the speech signal. Fig. 2 shows the waveform of the word 'soothing' and its short-time fractal dimension, average zero-crossing rate, and mean square amplitude as functions of time. While the fractal dimension D behaves similarly with zero-crossings, it has several advantages: For example, it can segment and distinguish between a vowel and a voiced fricative, whereas the zero-crossings usually fail because the rapid fluctuations of the voiced fricative may not appear as fluctuations around zero amplitude which would increase the zero-crossing rate but as a graph fragmentation which increases D. We have also observed cases where D could detect voiced stops but the zero-crossings could not.

Related to the Kolmogorov 5/3-law (3) is the fact that the variance between particle velocities at two spatial locations P and $P + \Delta P$ varies $\propto (\Delta P)^{2/3}$. These distributions have identical form to the case of fractional Brownian motion (Mandelbrot, 1982) whose variances scale with time differences T as T^{2H}, $0 < H < 1$, the frequency spectra vary $\propto 1/f^{2H+1}$ and time signals are fractal with dimension $D = 2 - H$. Thus, putting $H = 1/3$ leads to $D = 5/3$ for speech turbulence. Of course, Kolmogorov's law refers to wavenumber (not frequency) spectra and we dealt with pressure (not velocity) signals from the speech flow. Thus we should be cautious on how we interpret this result for speech. However, it is interesting to note that in our experiments with fricative sounds we observed D (for time scales < 0.1 ms) in the range $[1.65, 1.7]$ or often exactly 5/3=1.67. In previous work, Pickover and Khorasani (1986) reported global fractal dimensions of $D = 1.66$ for speech signals. However, they made no connection to the 5/3 law. Further, they used much longer time scales, i.e., 10 ms to 2 s and a different algorithm for computing fractal dimension. Thus, their work referred to time scales above the phoneme level, whereas our work is clearly below the phoneme time scale.

D Experiments Averaged over Multiple Phonemic Instances

In Fig. 3 the short-time fractal dimension D is shown computed over scales from 1/16 to 4 ms, using a 20 ms analysis window. For each phoneme the mean and standard deviation (shown as error bars) of the MFD is computed from 200 instances (100 from male and 100 from female speakers) of each phoneme in the TIMIT database. These experiments reinforce the claims made in the previous section that 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.

In Fig. 4(a), we compare the multiscale fractal dimension for the unvoiced fricative /sh/, the corresponding voiced fricative /zh/ and the vowel /uh/, averaged over 200 phonemic instances obtained from the TIMIT database. Clearly the small and medium scale fractal dimension measurement is smaller for voiced than for unvoiced sounds. The MFD is very small for vowels.

Plosives are a highly confusable set of phonemes. Multiscale fractal dimension 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/ etc. For example, in Fig. 4(b) we show the MFD for the voiced–unvoiced plosive pair /d/ and /t/ averaged over 200 occurrences. 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.

III Application to Automatic Speech Recognition

It has been demonstrated that the multiscale fractal dimension can potentially be used to discriminate among phonetic classes. 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 two-dimensional (2-D) distribution in time and scale. The main issue is how to represent this 2-D 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)[1] was arbitrarily chosen and is being used to compute the fractal features in the remainder of this paper. The 'standard' speech recognition features (i.e., cepstrum and mean square amplitude) are 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 Figs. 3 and 4 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. In this paper, we chose the simplistic 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 (Cole *et al*, 1990). The HMM-based HTK recognition package was used for all experiments (Young, 1995). A hold-one-out ("round-robin") procedure was used during training so that all 2700 words were available for testing. As a result the statistical significance of recognition performance comparisons was five times higher than in Singer and Lippmann (1992).

The 'standard' feature set consisted of the mean square amplitude (usually called 'energy'[2]) A, the first twelve cepstrum coefficients $C_1..C_{12}$ computed from a mel filterbank (Davis and Mermelstein, 1992) and their first time derivatives ΔA and $\Delta C_1..\Delta C_{12}$. The 'standard' feature vector was augmented by the fractal dimension of 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'.[3] Five-state left-right hidden Markov models were used in these experiments. As shown in Table 1, combining

[1] We chose to update all features every 10 ms because it is unclear how to incorporate features sampled with different frequencies in the HMM framework.

[2] We prefer the term 'mean square amplitude' over the term 'energy' because, as Kaiser (1990) has discussed, the energy in an oscillatory signal is more appropriate to be related to the physical energy of the source producing this signal. Such an energy is proportional both to the oscillation amplitude squared and the frequency squared and can be measured via the Teager-Kaiser energy operator.

[3] All stream weights are assumed to be unity.

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 $\{\Delta\Delta\mathrm{A}, \Delta\Delta\mathrm{C}_1..\Delta\Delta\mathrm{C}_{12}\}$ 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. Note that similar recognition performance (about 10-15% word error rate) was reported for the ISOLET database in Singer and Lippmann (1992).

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

IV Conclusions

In this paper, motivated by considerations based on the dynamics of the speech airflow, we have conjectured that short-time speech sounds contain various degrees of turbulence at multiple time scales below the phoneme time scale. To quantify these various degrees of turbulence we have proposed the use of the multiscale fractal dimension (MFD), measured via an efficient signal processing algorithm based on simple morphological dilation/erosion operators. Several experimental observations have been made by measuring the MFD of short-time speech sounds and demonstrating its potential for classification into certain broad phonetic classes and for speech segmentation.

Motivated by the novel information that the MFD can extract from the speech waveforms, we investigated the application of MFD as a 'fractal feature vector' to the problem of HMM-based automatic speech recognition. By augmenting the standard feature vector (containing short-time spectral information) used in current speech recognition systems with elements of the MFD vector

we have experimentally observed an improvement in performance, i.e., a modest reduction in the error of certain word recognition tasks over standard speech databases.

For future research, there are certain issues relating to the design of the classifier and the augmentation of the feature vector with fractal features that deserve further investigation. Such issues include the dimensionality of the fractal feature vector and the time scales of the fractal dimension used as features during recognition. The choice of the duration of the short-time analysis frame and alternative ways of incorporating the fractal feature vector in the HMM framework should also be considered more carefully. Additional performance improvements may be achieved with a more careful choice of these parameters. Further, despite the novelty of the information represented by the MFD vector, the improvement in performance that we observed when combining spectral and fractal features turned out to be relatively modest for HMM models with high complexity. This relatively small improvement in performance could be due to a correlation between the 'standard' spectral features and the multiscale fractal dimension. Specifically, preliminary experiments lead us to pose a question whether the fractal dimension is correlated with the high-frequency part of the spectrum. Thus, a formal study might be useful to investigate the existence and degree of any possible correlation between spectral features and multiscale fractal dimension.

Acknowledgments

We gratefully acknowledge the support of this research by the US National Science Foundation under Grants MIP-93963091 and MIP-9421677. We also wish to thank the anonymous reviewer for the many detailed comments and constructive suggestions, and especially for the observation that the estimated fractal dimension of a sampled sine increases with the sine frequency, all of which have helped us improve this paper.

References

Cole, R., Muthusamy, Y., and Fanty, M. (**1990**). "The ISOLET Spoken Letter Database," Tech. Rep. CSE 90-004, Oregon Graduate Institute of Science and Technology, Portland, Oregon.

Davis, S., and Mermelstein, P. (**1992**). "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, pp. 357–366.

Dubuc, B., Quiniou, J. F., Roques-Carmes, C., Tricot, C., and Zucker, S. W. (**1989**). "Evaluating the Fractal Dimension of Profiles," Phys. Rev. A, vol. 39, pp. 1500-1512.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, Hague).

Flanagan, J. L. (**1972**). *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, New York).

Kaiser, J. F. (**1983**). "Some Observations on Vocal Tract Operation from a Fluid Flow Point of View," in *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, I. R. Titze and R. C. Scherer (Eds.), The Denver Center for the Performing Arts, Denver, CO, pp. 358–386.

Kaiser, J. F. (**1990**). "On a Simple Algorithm to Calculate the 'Energy' of a Signal," in Proc. IEEE Int'l Conf. Acoust. Speech, and Signal Processing, Albuquerque, New Mexico, pp. 381–384.

Klatt, D. H. (**1987**). "Review of Text-To-Speech Conversion for English," J. Acoust. Soc. Am., vol. 82(3), pp. 737-793.

Mandelbrot, B. B. (**1982**). *The Fractal Geometry of Nature* (W.H. Freeman & Co., New York).

Maragos, P. (**1991**). "Fractal Aspects of Speech Signals: Dimension and Interpolation," Proc. IEEE Int'l Conf. Acoust., Speech, and Signal Processing, Toronto, Canada, pp. 417–420.

Maragos, P. (**1994**). "Fractal Signal Analysis Using Mathematical Morphology," in *Advances in Electronics and Electron Physics*, vol. 88, P. Hawkes and B. Kazan, Eds., Acad. Press.

Maragos, P., and Sun, F. K. (**1993**). "Measuring the Fractal Dimension of Signals: Morphological Covers and Iterative Optimization," IEEE Trans. Signal Processing, vol. 41, pp. 108–121.

McGowan, R. S. (**1988**). "An Aeroacoustics Approach to Phonation," J. Acoust. Soc. Am., vol. 83 (2), pp. 696-704.

Peitgen, O., Jürgens, H., and Saupe, D. (**1992**). *Chaos and Fractals* (Springer-Verlag, New York).

Pickover, C. A., and Khorasani, A. (**1986**). "Fractal Characterization of Speech Waveform Graphs," Comput. & Graphics, vol. 10, pp. 51-61.

Serra, J. (**1982**). *Image Analysis and Mathematical Morphology* (Acad. Press, New York).

Singer, E., and Lippmann, R. P. (**1992**). "A Speech Recognizer Using Radial Basis Function Neural Networks in an HMM Framework," Proc. IEEE Int'l Conf. Acoust., Speech, and Signal Processing, San Francisco, California, pp. 629–632.

Stevens, K. N. (**1971**). "Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations," J. Acoust. Soc. Am., vol. 50, no. 4 (2), pp. 1180-1192.

Teager, H.M., and Teager, S.M. (**1990**). "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," in *Speech Production and Speech Modeling*, W. J. Hardcastle and A. Marchal, Eds., NATO Advanced Study Institute Series D, Vol. 55, Bonas, France, July 1989; Kluwer Academic Publ., Boston, MA, pp. 241–261.

Thomas, T. J. (**1986**). "A Finite Element Model of Fluid Flow in the Vocal Tract," Comput. Speech & Language, vol. 1, pp. 131-151.

Tritton, D. J. (**1988**). *Physical Fluid Dynamics* (Oxford Univ. Press, New York).

Young, S. (**1995**). *The HTK Book* (Cambridge Research Lab: Entropics, Cambridge, England).

List of Figures

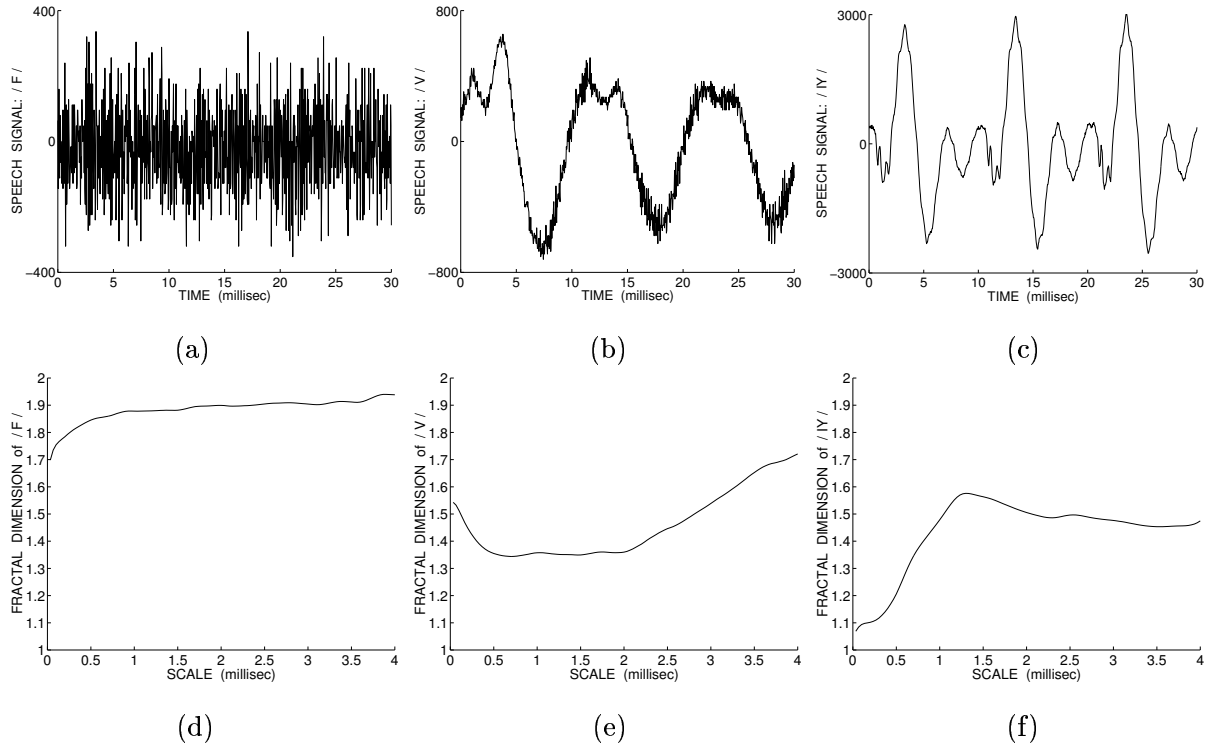

Figure 1: Top row shows waveforms from speech sounds sampled at 30 kHz. Bottom row shows their multiscale fractal dimensions estimated over moving windows of 10 scales.

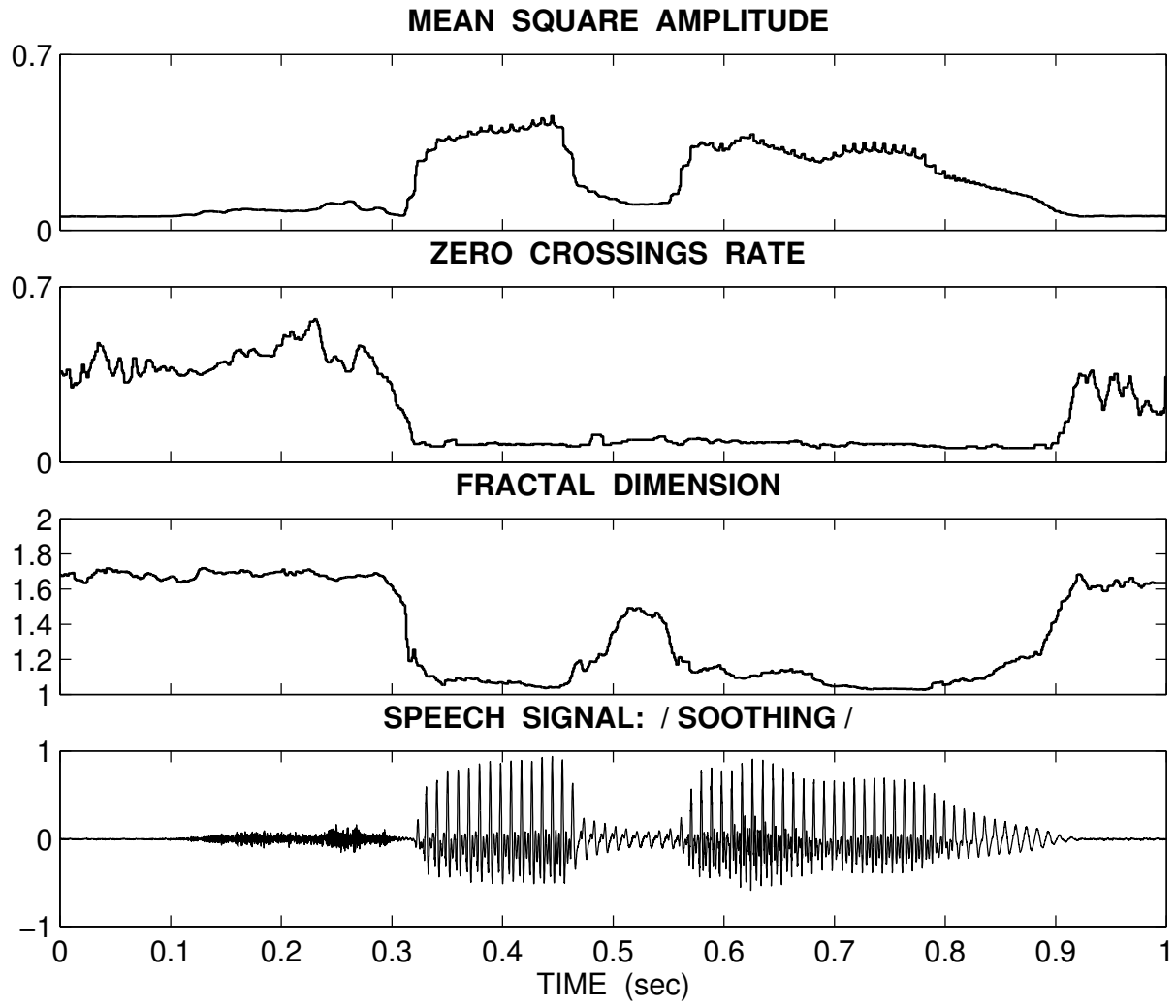

Figure 2: Speech waveform of the word 'soothing' sampled at 10 kHz and short-time speech measurements (fractal dimension at scale $\varepsilon = 1$, normalized zero-crossings rate, and normalized mean square amplitude) over a 10 ms window, computed every 1 ms and post-smoothed by a 3-point median filter.

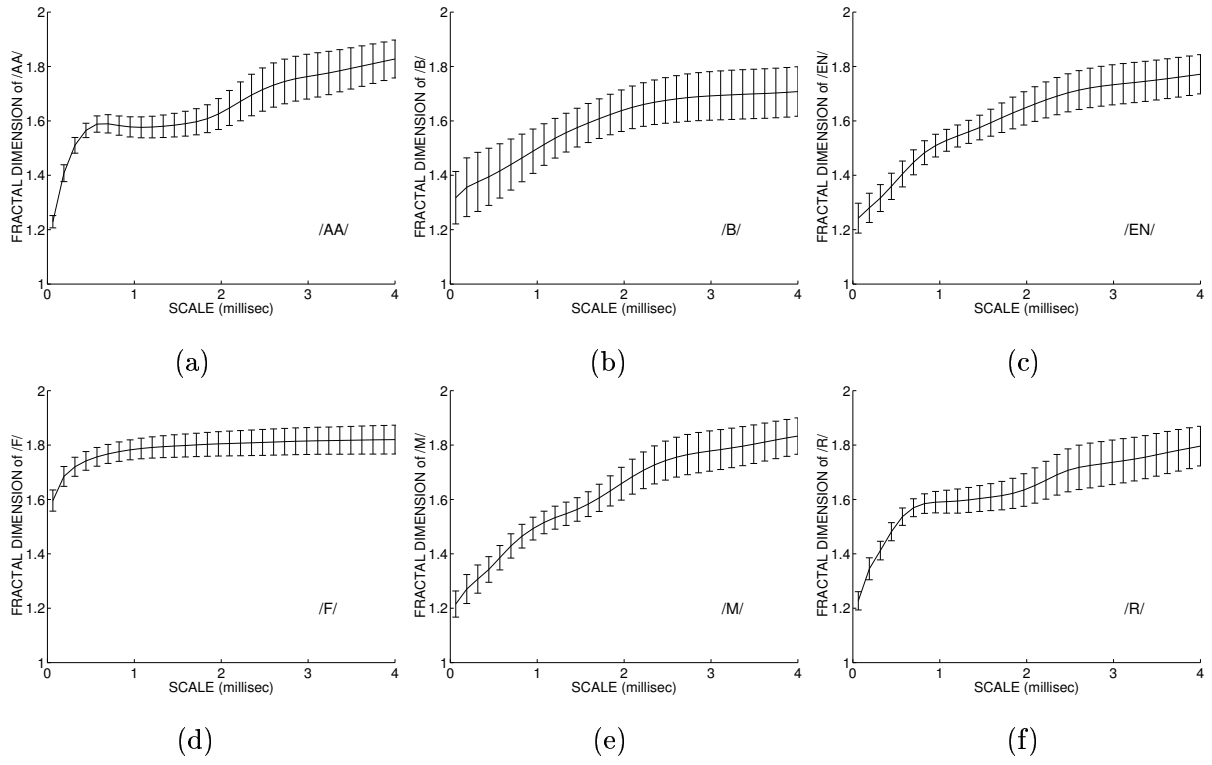

Figure 3: Mean and standard deviation (error bars) of the multiscale fractal dimension distribution for the phonemes /aa/, /b/, /en/, /f/, /m/, /r/ calculated from the TIMIT database (20 ms analysis window, updated every 10 ms).

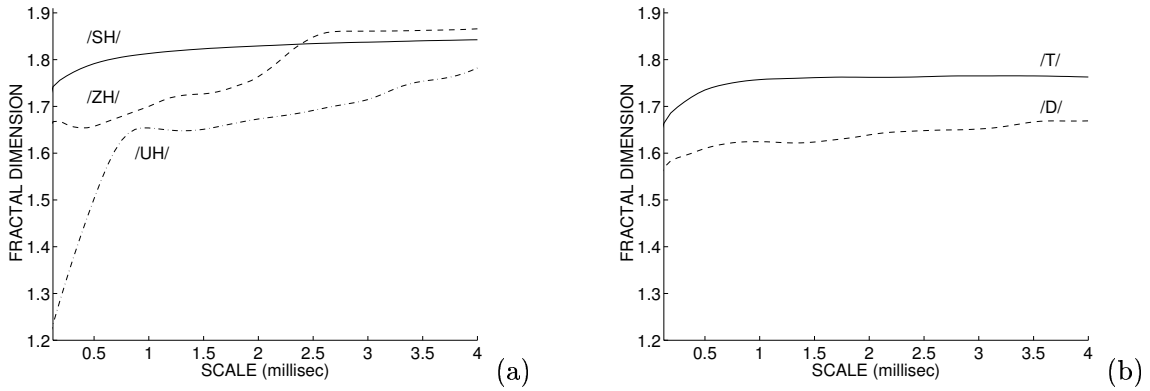

Figure 4: Multiscale fractal dimension: (a) phonemes /sh/, /zh/, /uh/ and (b) phonemes /t/, /d/; averaged over 200 phonemic instances from the TIMIT database (20 ms analysis window, updated every 10 ms).

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

$\{\mathrm{A}, \mathrm{C}_1..\mathrm{C}_{12}, \Delta\mathrm{A}, \Delta\mathrm{C}_1..\Delta\mathrm{C}_{12}\}$	$\{\mathrm{A}, \mathrm{C}_1..\mathrm{C}_{12}, \Delta\mathrm{A}, \Delta\mathrm{C}_1..\Delta\mathrm{C}_{12}\}$ + $\{D_1, \Delta D_1\}$	$\{\mathrm{A}, \mathrm{C}_1..\mathrm{C}_{12}, \Delta\mathrm{A}, \Delta\mathrm{C}_1..\Delta\mathrm{C}_{12}\}$ + $\{D_1, D_{11}, \Delta D_1, \Delta D_6, \Delta D_{11}, \Delta D_{16}\}$
81.2%	83.5 %	84.5%

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

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