FRACTAL ASPECTS OF SPEECH SIGNALS: DIMENSION AND INTERPOLATION

Petros Maragos

Division of Applied Sciences, Harvard University, Cambridge, MA 02138, USA

ABSTRACT: The nonlinear 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 and use it for speech segmentation and sound classification. We also develop a method for fractal speech interpolation, which can be used to synthesize controlled amounts of turbulence in speech or to increase its sampling rate by preserving not its bandwidth (as classically done) but rather its fractal dimension.

1 Speech Aerodynamics and Fractals

Preservation of momentum in the air flow during speech production yields the Navier-Stokes governing equation [8]

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

where ρ is the air density, p is the air pressure, $\vec{u}$ is the (vector) air particle velocity, and μ is the (assumed constant) air viscosity coefficient. Flow compressibility is assumed negligible since in speech flow (Mach numbers)$^2 \ll 1$. The Reynolds number $\mathrm{Re}=\rho U L/\mu$ characterizes the type of flow, 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 to have a much larger order of magnitude than the viscous forces. 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 flow region of similar (or constant) vorticity $\vec{\omega}$, where $\vec{\omega} = \nabla \times \vec{u}$. Vortices in the speech air flow have been experimentally found above the glottis in [6, 7] and theoretically predicted in [6, 4] 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 [6] 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, governed by the vorticity equation [8]

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

This research was supported in part by the Army Research Office under Grant DAALO3-86-K-0171 to the Center for Intelligent Control Systems, and in part by the National Science Foundation under Grant MIPS-86-58150 with matching funds from Bellcore, DEC, and Xerox.

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 into *turbulent flow*, which is a 'state of continuous instability' [8] 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 turbulence starts for jets at much lower Re than for flows attached to walls (as is the case downstream the vocal tract).

Modern theories that attempt to explain turbulence [8] 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 is transferred hierarchically to the small-size eddies which dissipate it due to viscosity. A related result is the famous Kolmogorov law

$$E(k,r) \propto r^{2/3} k^{-5/3} \quad (k \text{ in a finite range}) \qquad (3)$$

where $k = 2\pi/\lambda$ is the wavenumber, r is the energy dissipation rate, and $E(k,r)$ is the velocity wavenumber spectrum, i.e., Fourier transform of spatial correlations. In some cases this multiscale structure of turbulence can be quantified by *fractals*. Mandelbrot [2] 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. Several researchers also attempt to understand aspects of turbulence as cases of deterministic chaos. Chaotic dynamical systems converge to attractors whose sets in phase space or related time-series signals can be modeled by fractals. Now there are several mechanisms in high-Re speech flows that can be viewed as routes to chaos; e.g. vortices twist, strech, and fold (due to the bounded tract geometry). This process of twisting, stretching, and folding has been found in low-order nonlinear dynamical systems to give rise to chaos and fractal attractors.

All the above theoretical considerations and experimental evidence motivated in this paper our use of fractals as a mathematical and computational vehicle to analyze and synthesize 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 fragmentation. Since the relationship between turbulence and its fractal geometry or the fractal dimension of the resulting signals is currently very little 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 and segmentation. To measure it we developed a simple and efficient algorithm based on multi-

CH2977-7/91/0000-0417 $1.00 © 1991 IEEE

scale morphological filters that iteratively expand and contract the signal's graph. We also introduce a method for synthesizing controlled amounts of turbulence in speech by fractal interpolation: i.e., we increase the speech sampling rate by synthesizing fractal functions of prescribed fractal dimension that interpolate the original low-rate speech data. Interpolating speech is very important for multirate analysis and coding and has been classically done by preserving its bandwidth. In contrast, our speech interpolation method offers preservation of its fractal dimension.

2 Fractal Dimension of Speech

Let the continuous function $S(t)$, $0 \leq t \leq T$, represent a short-time speech signal, let the set $\mathcal{X} \subseteq \mathbf{R}^2$ represent its graph, and let D_H be the Hausdorff dimension of $\mathcal{X}$. The signal S is called *fractal* if its graph is a fractal set [2], i.e., if $D_H > 1$. Next we discuss two other dimensions closely related to D_H.

Minkowski-Bouligand dimension D_{MB}: Dilate $\mathcal{X}$ with a disk of radius ε and thus create a Minkowski *cover*. Find the area $A(\varepsilon)$ of the cover, and set its multiscale length equal to $\lim_{\varepsilon \to 0} L(\varepsilon)$, where $L(\varepsilon) = A(\varepsilon)/2\varepsilon$. Then D_{MB} is the constant D in the power law $L(\varepsilon) \propto \varepsilon^{1-D}$ as $\varepsilon \to 0$, which $L(\varepsilon)$ obeys if $\mathcal{X}$ 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 $\mathcal{X}$. Then $D_B = \lim_{\varepsilon \to 0} \log[N(\varepsilon)]/\log(1/\varepsilon)$ is obtained by replacing the Minkowski cover area with the box cover area.

In general, $1 \leq D_H \leq D_{MB} = D_B \leq 2$. In this work we focus only on D_{MB} which we shall henceforth call the "fractal dimension" D, because: 1) It coincides with D_H in many cases of practical interest; 2) It is much easier to compute than D_H. 3) It will be applied to sampled signals where most approaches can yield only approximate results. 4) It 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. ($D_B = D_{MB}$ in the continuous-time case, but they correspond to two different algorithms (with different performances) for sampled signals.)

As shown in [3], D will not change if we replace the disks in the Minkowski cover of $\mathcal{X}$ with other compact convex symmetric shapes $B \subseteq \mathbf{R}^2$. Thus, if $\varepsilon B = \{\varepsilon b : b \in B\}$ is an ε-scaled shape B, we can obtain multiscale multishape area distributions $A_B(\varepsilon) = \text{Area}(\mathcal{X} \oplus \varepsilon B)$, where $\oplus$ is the morphological dilation. Assuming now that $A_B(\varepsilon) \propto \varepsilon^{2-D}$ as $\varepsilon \to 0$ yields that

$$\log \frac{A_B(\varepsilon)}{\varepsilon^2} = D \cdot \log(\frac{1}{\varepsilon}) + \text{constant} \ , \ \text{as} \ \varepsilon \to 0. \qquad (4)$$

Instead of implementing the two-dimensional dilation $\mathcal{X} \oplus \varepsilon B$, it is computionally more efficient [3] to obtain $A_B(\varepsilon)$ via one-dimensional dilations $S \oplus G_\varepsilon$ and erosions $S \ominus G_\varepsilon$ of the signal S by the function $G_\varepsilon(t) = \sup\{y \in \mathbf{R} : (t, y) \in \varepsilon B\}$ at all scales ε:

$$A_B(\varepsilon) = \int_0^T [S \oplus G_\varepsilon(t) - S \ominus G_\varepsilon(t)]dt. \qquad (5)$$

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.

For a discrete-time finite speech signal $S[n]$, $n = 0, 1, \ldots, N$, we use covers at discrete scales $\varepsilon = 1, 2, \ldots$, and restrict the discrete set B to have radius=1. Then the corresponding 3-sample function $G[n]$ (at scale $\varepsilon = 1$) can have only two shapes: a triangle or a rectangle [3]. In both cases $G[0] = h \geq 0$ is allowed to vary and match the amplitude range of the signal S. The main result in the discrete case is the following recursive algorithm [3]:

$$\begin{aligned} S \oplus G[n] &= \max_{-1 \leq k \leq 1} \{S[n+k] + G[k]\} \ , \quad \varepsilon = 1 \\ S \oplus G_{(\varepsilon+1)} &= (S \oplus G_\varepsilon) \oplus G \ , \qquad \varepsilon = 2, \ldots, \varepsilon_{max} \end{aligned} \qquad (6)$$

Likewise for the erosions $S \ominus G_\varepsilon$. Next, we compute the areas $A_B[\varepsilon]$ by replacing the $\int_0^T$ in (5) with summation $\sum_{n=0}^{N}$. Finally, we fit a straight line using least-squares to the plot of $(\log A_B[\varepsilon]/\varepsilon^2, \log 1/\varepsilon)$, whose slope gives the fractal dimension of S. For "real world" fractal signals the assumption of a constant D at all scales ε is not true. Hence, instead of a global dimension, we estimate the *local fractal dimension* LFD[ε], which for each ε is equal to the slope of a line fitted to the log-log plot of (4) over a moving window $\{\varepsilon, \varepsilon + 1, \ldots, \varepsilon + 9\}$ of 10 scales. We henceforth select $h = 0$, which makes G a binary function, because then the erosions/dilations can be performed faster, and the algorithm yields fractal dimensions that are invariant to any affine transformation $S \mapsto aS + b$ of the amplitude range ($a > 0$).

Figure 1 shows 30 msec 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 LFD[ε] for $\varepsilon = 1, .., 90$, i.e., for time scales 1/15 - 6 msec. The reason for the higher than usual sampling rate is to preserve the fragmentation of the sampled signal as close as possible to that of the continuous-time speech signal. We have conducted many experiments similar to Fig. 1, from which we concluded the following: 1) Unvoiced fricatives (/F/, /θ/, /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 msec), consistent with the turbulence phenomena present during their production. 2) Vowels at small scales (< 0.1 msec) 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 $> 2 - 3$ msec (i.e., at scales of the same order as the distance between their major consecutive peaks) their dimension increases appreciably. 3) Some voiced fricatives like /V/ and /TH/ have a mixed behavior. If they don't contain a fully developed turbulence state, at scales < 0.1 msec they have a medium fractal dimension $D \in [1.3, 1.6]$, which increases at scales > 5 msec (for the same reasons as for vowels) and may decrease for intermediate scales. Otherwise, their dimension is high (> 1.6), although often somewhat lower than that of their unvoiced counterparts. Thus, for normal conversational speech, we have found that its short-time (e.g., over $\sim 10 - 30$ msec frames) fractal dimension D (evaluated at a scale < 0.1 msec) can roughly distinguish these three broad classes of speech sounds by quantifying the amount of their waveform's fragmentation. However, for loud speech (where the air velocity and Re increase, and hence the onset of turbulence is easier) or for breathy voice (especially for female speakers) the dimension of several speech sounds, e.g. vowels may significantly increase. In general, the D 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 D since some signal's fragmentation has been lost during sampling), and c) the speaking state. Therefore, we often don't assign any particular importance to the absolute D estimates but only to their average ranges and relative differences.

We also used D estimated at a single small time scale, i.e., LFD[$\varepsilon = 1$], 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 a word and its short-time fractal dimension, average zero-crossing rate, and energy as functions of time. While 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 can fail (see Fig. 2) because the rapid fluctuations of the voiced fricative may not appear as zero-mean oscillations 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 of velocity differences between two points at distance ΔX varies $\propto (\Delta X)^{2/3}$. These distributions have identical form to the case of fractional Brownian motions [2] whose variances scale with time differences ΔT as $(\Delta T)^{2H}$, $0 < H < 1$, 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 often observed D (for time scales < 0.1 msec) in the range $[1.65, 1.7]$ or sometimes exactly 5/3=1.67. In [5] a global dimension $D = 1.66$ was reported for speech signals but no mention of the 5/3 law was made, the D estimation algorithm was different, and the time scales were much longer, i.e., 10 msec to 2 sec; thus in [5] the time scales were above the phoneme level, whereas our work is clearly below the phoneme time scale.

3 Fractal Interpolation of Speech

Bandlimited discrete signal interpolation has been done traditionally by upsampling (i.e., by inserting zeros between consecutive signal samples) and passing the upsampled signal through a low-pass linear filter to smooth the abrupt transitions during gaps, while preserving the bandwidth. Given the importance of fractal dimension for speech, we develop here an alternative approach to interpolate speech by synthesizing and upsampling a fractal function that interpolates the given low-rate speech and can have any desired fractal dimension.

Before we discuss fractal speech interpolation, we summarize basic ideas from the theory of fractal interpolation functions [1]. Given is a set of data points $\{(x_n, y_n) \in \mathbf{R}^2; n = 0, 1, 2, ..., N > 1\}$ on the plane, where $x_{n-1} < x_n$. In the complete metric space $\mathcal{G}$ of all continuous functions $g : [x_0, x_N] \to \mathbf{R}$ such that $g(x_0) = y_0$ and $g(x_N) = y_N$ define the function mapping Ψ by

$$\Psi(g)(x) = c_n(x - b_n)/a_n + R_n g((x - b_n)/a_n) + d_n \quad , \quad x \in [x_{n-1}, x_n] \tag{7}$$

where $n = 1, 2, ..., N$, the $R_n \in (-1, 1)$ are free parameters, and the $4N$ parameters a_n, b_n, c_n, d_n are uniquely determined by

$$a_n x_0 + b_n = x_{n-1} \quad , \quad a_n x_N + b_n = x_n \tag{8}$$

$$R_n y_0 + c_n x_0 + d_n = y_{n-1} \quad , \quad R_n y_N + c_n x_N + d_n = y_n \tag{9}$$

Under the action of Ψ the graph of the input function g is mapped to the graph of the output $\Psi(g)$ via affine mappings $(x, y) \mapsto (ax + b, Ry + cx + d)$, which include contractions and shifts of the domain and range of g. Ψ is a contraction mapping in $\mathcal{G}$ and has a unique fixed point which is a continuous function $F : [x_0, x_N] \to \mathbf{R}$ that interpolates the given data; i.e., $F(x_n) = y_n$ for $n = 0, 1, ..., N$. F is called a *fractal interpolation function*, because quite often the fractal dimension D of its graph $\mathcal{X}$ exceeds 1. Specifically, if $\sum_{n=1}^{N} |R_n| > 1$ and (x_n, y_n) are not all collinear, then D is the unique real solution of $\sum_{n=1}^{N} |R_n| a_n^{D-1} = 1$; otherwise, $D = 1$. If $a_n = 1/N$ for all n, then

$$D = 1 + \frac{\log\left(\sum_{n=1}^{N} |R_n|\right)}{\log N} \tag{10}$$

Thus by choosing the vertical scaling ratios R_n's we can synthesize a fractal interpolation function of any desired fractal dimension. F can be synthesized by iterating Ψ on any initial fuction g in $\mathcal{G}$; i.e., $F = \lim_{k \to \infty} \Psi^{ok}(g)$ where $\Psi^{ok}(g) = \Psi[\Psi^{o(k-1)}(g)]$.

Let $S_o[n]$, $n = 0, 1, .., N$ be an original short-time speech segment. To fractally interpolate it by a factor L we start from the $N + 1$ data pairs $(x_n = nL, y_n = S_o[n])$ with $x_N = NL = M$, set $a_n = 1/N$, $b_n = x_{n-1}$, and select $R_n = R \in (-1, 1)$. Then there is a unique fractal interpolation function $F(x)$, $x \in [0, M]$, which interpolates the given data, i.e, $F(nL) = S_o[n]$. If $R = 0$, F is the piecewise-linear interpolant of the data. The graph of F has fractal dimension $D = 2 + \log|R| / \log N$ if $1 > |R| > 1/N$, and $D = 1$ if $|R| \leq 1/N$. The larger $|R|$, the larger D, the rougher F looks. Based on F we can up-sample S_o to a $1 : L$ interpolated signal $S_i[m] = F(m)$, $m = 0, 1, .., M$. We use the single scaling ratio R as a parameter to control the fractal dimension (and hence the amount of turbulence) in the interpolated speech. If we select $R = N^{D-2}$, where D is the measured short-time fractal dimension of the low-rate signal S_o, then we essentially interpolate speech by preserving its fractal dimension. Fig. 3a shows a 26.7 msec original speech segment from the voiced fricative /V/ sampled at 30 KHz. This was decimated to a signal S_o at 6 KHz; Fig. 3b shows S_o upsampled at 30 KHz using the classical bandlimited 1:5 interpolation. Fig. 3c shows the 1:5 fractal interpolation S_i of S_o using a ratio R corresponding to a measured $D = 1.66$. We see that the bandlimited interpolation cannot reconstruct some of the high frequency structure since it preserves the original (3 KHz) bandwidth, whereas the fractal interpolation (preserving the signal's fractal dimension) reconstructs part of this high frequency structure. We have also applied this fractal interpolation method to speech unvoiced fricatives and vowels and observed the same good consistency in the amount of fragmentation between the original high-rate speech and the fractally interpolated low-rate speech.

Acknowledgement: I wish to thank Howard Stone at Harvard for many useful discussions on fluid dynamics.

References

[1] M. Barnsley, *Fractals Everywhere*, NY: Acad. Press, 1988.

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

[3] P. Maragos & F. K. Sun, "Measuring the Fractal Dimension of Signals", Tech. Rep. CICS-P-193, Brown-Harvard-MIT Center for Intelligent Control Systems, Feb. 1990.

[4] R. S. McGowan, "An aeroacoustics approach to phonation", *J. Acoust. Soc. Am.*, 83 (2), pp.696-704, Feb. 1988.

[5] C. Pickover & A. Khorasani, "Fractal Characterization of Speech Waveform Graphs", *Comp. & Graphics*, 10, 1986.

[6] H.M. Teager & S.M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", *Proc. NATO ASI on Speech Production and Speech Modelling*, France, 1989.

[7] T. J. Thomas, "A finite element model of fluid flow in the vocal tract", *Comput. Speech & Language*, 1:131-151, 1986.

[8] D. J. Tritton, *Physical Fluid Dynamics*, Oxf. U. Pr., 1988.

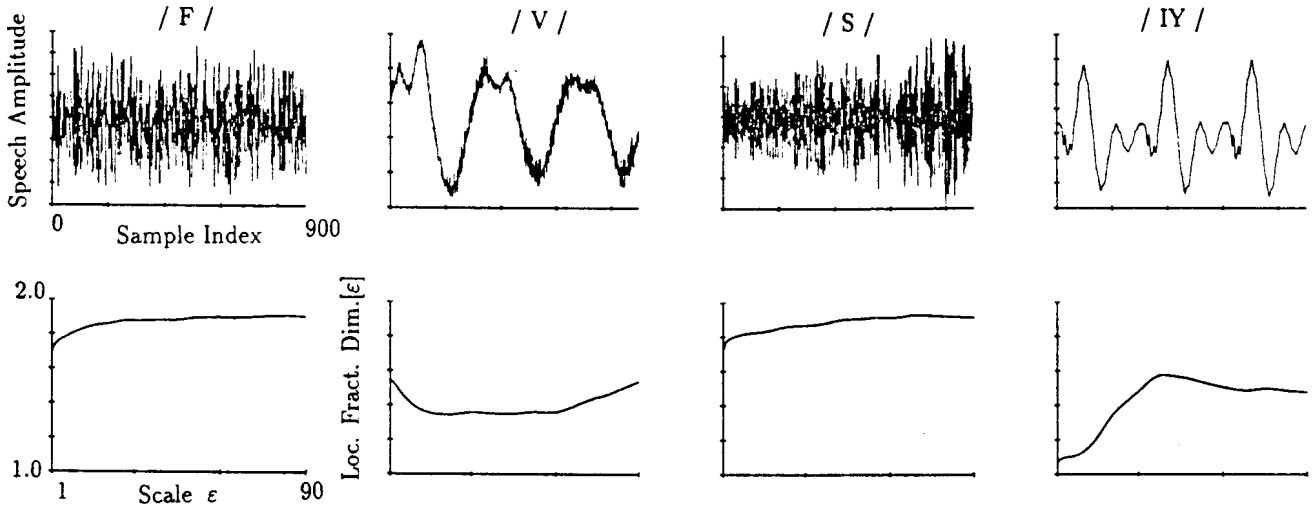

Figure 1. Top row shows waveforms from speech sounds. Bottom row shows their local fractal dimensions.

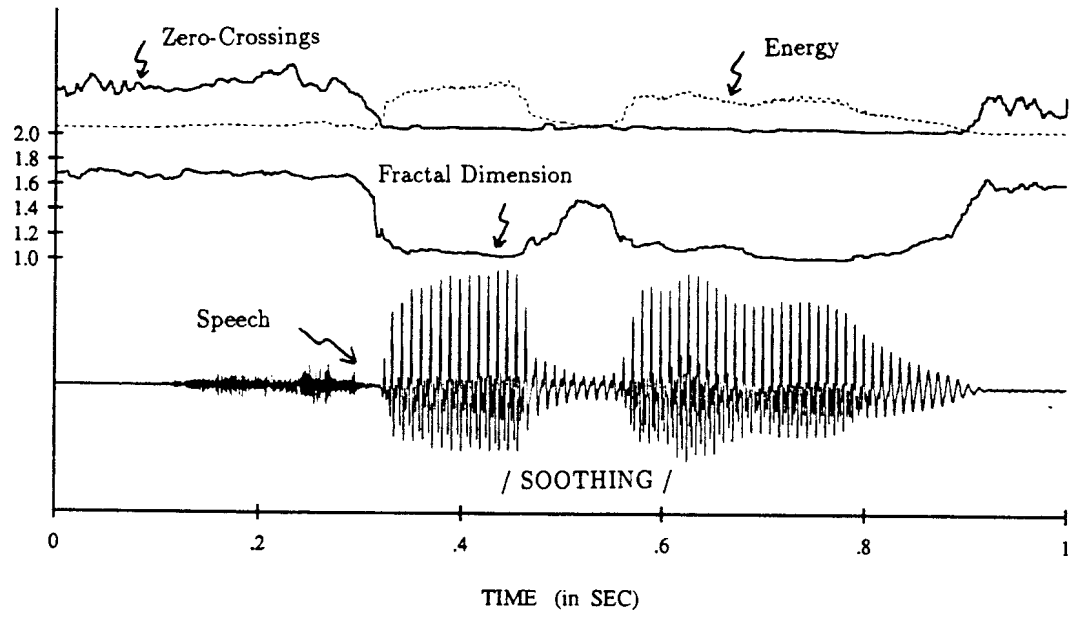

Figure 2. Speech waveform of the word /soothing/ sampled at 30 KHz and short-time speech measurements over a 10 msec window, computed every 1 msec and post-smoothed by a 3-point median filter.

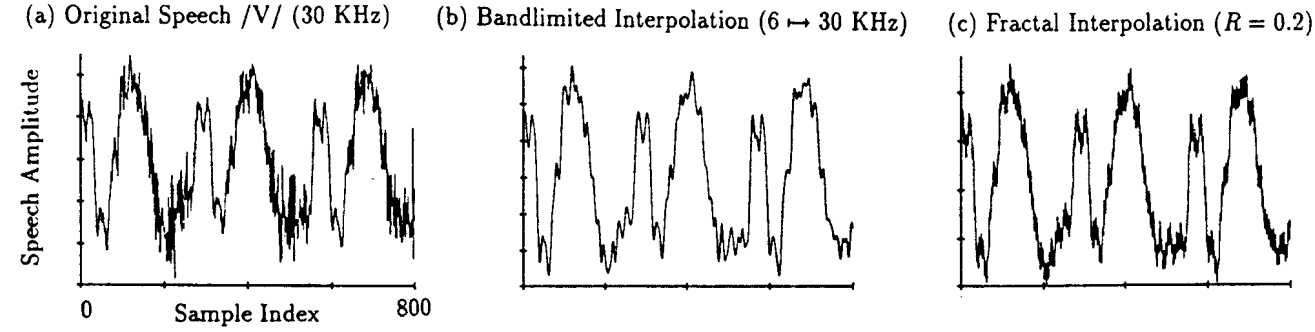

Figure 3. Fractal Interpolation of Speech ($N = 160$, $L = 5$, $M = 800$; $D = 1.66 \Longrightarrow R = 0.2$).