Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 1206–1223
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Analysis and classification of speech signals by generalized fractal dimension features

Vassilis Pitsikalis *, Petros Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytexneiou Str., Athens 15773, Greece

Received 7 August 2008; received in revised form 17 March 2009; accepted 1 June 2009

Abstract

We explore nonlinear signal processing methods inspired by dynamical systems and fractal theory in order to analyze and characterize speech sounds. A speech signal is at first embedded in a multidimensional phase-space and further employed for the estimation of measurements related to the fractal dimensions. Our goals are to compute these raw measurements in the practical cases of speech signals, to further utilize them for the extraction of simple descriptive features and to address issues on the efficacy of the proposed features to characterize speech sounds. We observe that distinct feature vector elements obtain values or show statistical trends that on average depend on general characteristics such as the voicing, the manner and the place of articulation of broad phoneme classes. Moreover the way that the statistical parameters of the features are altered as an effect of the variation of phonetic characteristics seem to follow some roughly formed patterns. We also discuss some qualitative aspects concerning the linear phoneme-wise correlation between the fractal features and the commonly employed mel-frequency cepstral coefficients (MFCCs) demonstrating phonetic cases of maximal and minimal correlation. In the same context we also investigate the fractal features' spectral content, in terms of the most and least correlated components with the MFCC. Further the proposed methods are examined under the light of indicative phoneme classification experiments. These quantify the efficacy of the features to characterize broad classes of speech sounds. The results are shown to be comparable for some classification scenarios with the corresponding ones of the MFCC features.

Keywords: Feature extraction; Generalized fractal dimensions; Broad class phoneme classification

1. Introduction

Well-known features, such as the mel-frequency cepstral coefficients (MFCCs), are based on the linear source-filter model of speech. This modeling approach when fertilized by auditory concepts that are incorporated via the mel-scale (Davis and Mermelstein, 1980) spacing of the filter-bank, results in a feature space representation that captures characteristics of the speech production system. Such feature space representations are massively utilized in automatic speech recognition (ASR) systems, which still suffer as far as plain acoustic modeling is considered. Herein, we investigate whether an alternative feature space representation that is taking advantage of a different perspective may be utilized for analysis, and furthermore for characterization of speech signals. Specifically, we exploit novel feature descriptions that are based on simple concepts from the *system dynamics* and *fractal theory*. Via the proposed analysis we seek to investigate the capability of the methods concerning speech sound characterization and relate general phonetic characteristics with the proposed measurements. A practical motivation that prompts this direction is the successful, to a certain degree, application of related methods in ASR (Maragos and Potamianos, 1999; Pitsikalis and Maragos, 2006). Hence, also continuing previous work, we focus on fractal features as these are related to a set of generalized fractal dimension measurements and furthermore proceed by considering

* Corresponding author. Tel.: +302 107722964; fax: +302 107723397.
E-mail addresses: vpitsik@cs.ntua.gr, vassilis.pitsikalis@gmail.com (V. Pitsikalis), maragos@cs.ntua.gr (P. Maragos).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.06.005

the following aspects: (1) provide statistical measurements regarding the fractal-related features, and discuss issues on the characterization of speech sounds by the new measurements; (2) apply a variety of classification experiments; and (3) highlight viewpoints concerning their correlation with the cepstral features.

The employed concepts from system dynamics and fractal theory originate from the experimental and theoretical evidence on the existence of nonlinear aerodynamic phenomena in the vocal tract during speech production (Teager and Teager, 1989; Kaiser, 1983; Thomas, 1986), such as flow separation, generation of vortices e.g. at the separation boundary, jet formation and its subsequent attachment to the walls; these phenomena, together with the possible generation of turbulent flow, indicate the nonlinear character of the speech production system also leading to a discussion on the factor by which they affect phonation (Hirschberg, 1992; Howe and McGowan, 2005). From the observation point of view, the dynamics of systems that demonstrate phenomena sharing characteristics with turbulent flow are referred to as 'chaotic' (Tritton, 1988; Peitgen et al., 1992). Such systems are characterized by limited predictability, whereas nonlinearity can be an essential feature of the flow. Turbulent motion can be seen as a combination of interacting motions at various length scales leading to the formation of 'eddies' (Tritton, 1988), i.e. localized structures of different sizes. Such structures function for the transfer of energy from higher to lower scales, until the extent of energy dissipation due to viscosity; a phenomenon known as the energy cascade. The twisting, stretching and folding that are accounted in this context are also characteristics of deterministic systems that resemble chaotic behaviour; these are characterized by properties such as mixing and conditional dependence on the initial conditions (Peitgen et al., 1992). Within this frame of reference, fractal dimensions and Lyapunov exponents are among the invariant quantities that may be used for the characterization of a chaotic system. Besides, it has been conjectured that methods developed in the frame of chaotic dynamical systems and fractal theory may be employed for the analysis of turbulent flow: for instance by utilizing fractals and multifractals to model the geometrical structures in turbulence that are related to phenomena such as the energy cascade (Mandelbrot, 1982; Benzi et al., 1984; Meneveau and Sreenivasan, 1991; Takens, 1981; Hentschel and Procaccia, 1983). For further discussion on this motivation see (Maragos and Potamianos, 1999). In general, fractal dimensions can be utilized to quantify the complexity, concerning the geometry of a dynamical system given its multidimensional phase-space. This quantification is related to the active degrees of freedom of the assumed dynamical system, providing a quantitative characterization of a system's state.

Recently there have been directions in speech analysis that are based on concepts of fractal theory and dynamical systems. Numerous methods have been proposed (Maragos et al., 1993; Narayanan and Alwan, 1995; Kumar and Mullick, 1996; Banbrook et al., 1999) that attempt to exploit the turbulence-related phenomena of the speech production system in some way . Work in this area includes the application of fractal-measures on the analysis of speech signals (Maragos, 1991; Maragos and Potamianos, 1999), application of nonlinear oscillator models to speech modeling, prediction and synthesis (Quatieri and Hofstetter, 1990; Townshend, 1991; Kubin, 1996), or multifractal aspects (Adeyemi and Boudreaux-Bartels, 1997). For instance (Maragos, 1991; Maragos and Potamianos, 1999), fractal dimensions are computed as an approximate quantitative characteristic that corresponds to the amount of turbulence that may reside in a speech waveform during its production, via the speech waveform graph's fragmentation. Ideas concerning phase-space reconstruction have attracted additional interest. Methods that follow this approach are based on the embedding theorem (Sauer et al., 1991). The analysis may be followed by measurement of invariant quantities on the reconstructed space. Early works in the field employing phase-space reconstruction include (Quatieri and Hofstetter, 1990; Townshend, 1991; Bernhard and Kubin, 1991; Herzel et al., 1993; Narayanan and Alwan, 1995; Kumar and Mullick, 1996; Greenwood, 1997), whereas recently there has been increasing interest in the area (Banbrook et al., 1999; Kokkinos and Maragos, 2005; Johnson et al., 2005). These employ concepts on Lyapunov exponents (Kumar and Mullick, 1996; Banbrook et al., 1999; Kokkinos and Maragos, 2005), density models of the phase-space (Johnson et al., 2005), correlation dimension measurements (Kumar and Mullick, 1996; Greenwood, 1997), especially for fricative consonants (Narayanan and Alwan, 1995), or surrogate analysis on the nonlinear dynamics of vowels (Tokuda et al., 2001).

In this paper, a speech signal segment is thought of as a 1-D projection of the assumed unknown phase-space of the speech production system. We reconstruct a *multidimensional* phase-space (Section 2) and aim to capture measures of the assumed speech production system's dynamics in the way that these are described by the reconstructed space. Such measures are related in our case to the fractal dimensions. Moreover the analysis with generalized fractal dimensions renders the detection of a set's inhomogeneity feasible.

Thus, as an extension of previous work (Maragos, 1991; Maragos and Potamianos, 1999), which exploits multiscale fractal dimension on the scalar 1-D speech signal, we move a step forward (Pitsikalis and Maragos, 2002; Pitsikalis et al., 2003), according to the directions outlined above employing measurements such as, the correlation dimension (Section 3.1) and especially the generalized dimensions (Section 3.2) on embedded spaces for the analysis of speech phonemes. Since related methods have been employed to a certain extent successfully in speech recognition applications (Maragos and Potamianos, 1999; Pitsikalis and Maragos, 2006), we take a closer look at the employed methods in the following ways. At first we highlight issues on their application in the practical cases of speech phonemes and

then construct simple descriptive feature vectors incorporating information carried by the raw measurements. We further demonstrate (Section 3.3) indicatively the statistical trends and patterns that the feature elements follow depending upon general properties such as the voicing, the manner and the place of articulation. In the same framework we show how phonetic properties affect statistical quantities related to the fractal dimensions (Section 3.4). An implicit indication on the relation between the information carried by the proposed fractal features and the commonly used MFCC is presented by measuring their in between correlation (Section 4). This lets us consider some novel viewpoints at first on how the correlation between the two feature sets varies with respect to the phoneme class, and secondly on the fractal features' spectral content as this is formed in maximal and minimal correlation cases with the MFCC. The potential of the measurements to characterize speech sounds is also investigated in the light of classification experiments that complement the preceding analysis (Section 5). These contain: (1) experiment sets of single phoneme classification tests; after inspecting characteristics on the phoneme confusability of the features, we proceed by considering, and (2) experiments on broad phoneme classes; in this way we examine quantitatively the efficacy of the proposed analysis from the viewpoint of the resulting discriminative ability. The fractal classification accuracies are also compared with two variants of MFCC-based baselines, showing in some classification scenarios comparable performance for the broad class case.

2. Embedding speech signals

We assume that in discrete time n the speech production system may be viewed as a nonlinear, but finite dimensional due to dissipativity (Temam, 1993), dynamical system $Y(n) \to F[Y(n)] = Y(n+1)$. A speech signal segment $s(n)$, $n = 1, \ldots, N$, can be considered as a 1-D projection of a vector function applied to an unknown *multidimensional* state vector $Y(n)$. Next, we employ a procedure by which a phase-space of $X(n)$ is reconstructed satisfying the requirement to be diffeomorphic to the original $Y(n)$ phase-space so that determinism and differential structure of the dynamical system are preserved. The embedding theorem (Packard et al., 1980; Takens, 1981; Sauer et al., 1991) provides the supporting justification to proceed while satisfying these requirements.

According to the *embedding* theorem (Sauer et al., 1991), the vector

$$X(n) = [s(n), s(n+T_D), \ldots, s(n+(D_E-1)T_D)] \tag{1}$$

formed by samples of the original signal and delayed by multiples of a constant time delay T_D defines a motion in a reconstructed D_E-dimensional space that shares common aspects with the original phase-space of $Y(n)$. Particularly, invariant quantities of the assumed dynamical system such as the fractal dimensions from $Y(n)$ are conserved in the reconstructed space traced by $X(n)$. Thus, by studying the constructible dynamical system $X(n) \to X(n+1)$ we can uncover useful information on the complexity as it is related to these invariant quantities about the original unknown dynamical system $Y(n) \to Y(n+1)$. The above is feasible provided that the unfolding of the dynamics is successful, e.g. the embedding dimension D_E is large enough. For instance, let us consider a toy system case where the original phase-space is known: if one uses smaller embedding dimension than the one required, the resulting reconstruction would suffer from collapsing points; these points would otherwise belong to separate time orbits. This would imply also a case of ambiguous determinism since there would be multiple possible dynamic orbits for the succeeding points in the time instances that follow. For further discussion on these issues see (Sauer et al., 1991). However, the embedding theorem does not specify any methods to determine the required parameters (T_D, D_E) but only sets constraints on their values. For example, D_E must be greater than twice the box-counting dimension of the multidimensional set.

The smaller the T_D gets, the more correlated shall the successive elements be. Consequently the reconstructed vectors will populate along the separatrix of the multidimensional space. On the contrary, the greater the T_D gets, the more random will the successive elements be and any preexisting order shall vanish. To compromise, the average mutual information I for the signal $s(n)$ is first estimated as

$$I(T) = \sum_{n=1}^{N-T} P(s(n), s(n+T)) \cdot \log_2 \left[\frac{P(s(n), s(n+T))}{P(s(n)) \cdot P(s(n+T))} \right] \tag{2}$$

where $P(\cdot)$ is a probability density function estimated from the histogram of $s(n)$. $I(T)$ is a measure of nonlinear correlation between pairs of samples of the signal segment that are T positions apart. Then, the time delay T_D is selected as

$$T_D = \min\{\arg\min_{T \geqslant s0} I(T)\} \tag{3}$$

The final step in the embedding procedure is to set the dimension D_E. As a consequence of the projection, points of the 1-D signal are not necessarily in their relative positions because of the true dynamics of the multidimensional system, referred to as true neighbors; manifolds are folded and different distinct orbits of the dynamics may intersect. A true versus false neighbor criterion is formed by comparing the distance between two points S_n, S_j embedded in successive increasing dimensions. If their distance $d_D(S_n, S_j)$ in dimension D is significantly different, for example by one order of magnitude, from their distance in dimension $D+1$, i.e.

$$R^D(S_n, S_j) = \frac{d_{D+1}(S_n, S_j) - d_D(S_n, S_j)}{d_D(S_n, S_j)} \tag{4}$$

exceeds a threshold (in the range 10–15) then they are considered to be a pair of *false neighbors*. Note that any difference in distance should not be greater than some second

order magnitude multiple of the multidimensional set radius $R_A = \frac{1}{N}\sum_{n=1}^{N}\|s(n) - \bar{s}\|$. The dimension D at which the percentage of false neighbors goes to zero, or is minimized in the existence of noise, is chosen as the embedding dimension D_E. An extensive review of such methods can be found in (Abarbanel, 1996; Kantz and Schreiber, 1997).

Following the procedures described we set the embedding parameters for the cases of speech signals and next construct the embeddings of three indicative types of phonemes. Fig. 1 illustrates a few multidimensional phonemes together with their corresponding scalar waveforms. Before the analysis and measurements of the following sections, it seems, by inspection of the multidimensional signals, that the different phoneme types are characterized in the reconstructed phase-spaces by different geometrical properties. For instance the vowel phoneme /ah/ demonstrates dynamic cycles that resemble laminar "flow" in the phase-space, the unvoiced fricative /s/ is characterized by many discontinuous trajectories, and the unvoices stop /p/ shows a single trajectory that settles to a region of interwoven tracks. Similar observations have been made since (Bernhard and Kubin, 1991; Herzel et al., 1993). Our goal is to describe this variation by means of statistical measurements that are related to the fractal dimensions.

3. Fractal dimensions and feature extraction

The Renyi hierarchy of generalized dimensions D_q, $q \geqslant 0$ is defined (Hentschel and Procaccia, 1983; Peitgen et al., 1992) by exploiting the exponential dependency, with respect to the order parameter q, of a set's natural measure. In this way it constructs a sequence that unifies and extends known fractal dimensions. Such cases are of geometrical type such as the box-counting dimension D_B corresponding to $q = 0$, or of probabilistic type such as the information D_I and correlation dimension D_C for $q = 1$ and $q = 2$, respectively. Our exploration of methods for the analysis of speech signals by fractal measurements has started (Maragos, 1991; Maragos and Potamianos, 1999) with the already presented multiscale fractal dimension (MFD) which corresponds to the D_B. In the sections that follows, a step ahead of these *first order* measurements employed on the scalar 1-D speech signal involves the exploitation of the multidimensional embedded speech signals. Towards speech feature extraction we consider at first measurements that are related to the correlation dimension D_C and a set of generalized dimensions that has been shown to extend the aforementioned cases of the Renyi set of fractal dimensions (Badii and Politi, 1985).

3.1. Correlation dimension

3.1.1. Background

The *correlation dimension* can be estimated by employing a practical method from the category of average point-wise mass algorithms for dimension estimation (Grassberger and Procaccia, 1983). A quantity used for its estimation is the correlation sum C that measures how often a typical sequence of points visits different regions of the set and quantifies its mass in this way. C is given for each scale r by the number of points with distances less than r normalized by the number of pairs of points:

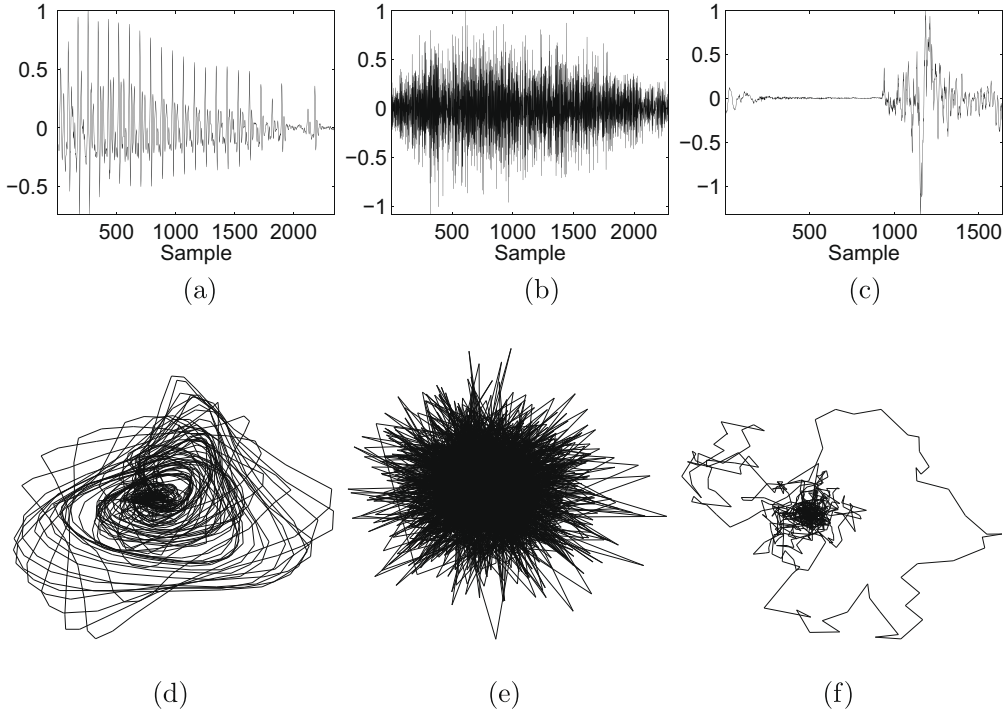


Fig. 1. Phoneme signals from the TIMIT database (upper row) with the corresponding embeddings (bottom row): (a,d) /ah/, (b,e) /s/, (c,f)/p/.

$$C(N,r) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \theta(r - \|X_i - X_j\|) \quad (5)$$

where θ is the Heaviside unit-step function. The correlation dimension is then defined as

$$D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C(N,r)}{\log r} \quad (6)$$

For small enough scales and for large enough N, $C(r)$ is proportional to r^{D_C}.

3.1.2. CD features and characterization of speech signals

In the *unfolded* phase-space we measure C and D_C as in (6) using least squares local slope estimation of the $\log C(N,r)$ versus $\log r$ data, weighted with the corresponding variance of each set of points. In this way we form the local-scale correlation dimension function $D_C(r)$ with respect to the local-scale parameter $r \in [r_{min}, r_{max}]$. The scale boundaries are selected by ignoring a small percentage of scales at each extent (Kantz and Schreiber, 1997). In order to derive information from the set of raw measurements we form the following 8-dimensional feature vector, whose elements are related to the correlation dimension (CD). This concerns both the sum C, i.e. the average pairwise correlation over the whole set, and how this quantity is varying in terms of the scale parameter's exponent $D_C(r)$. The feature components represent the measurements by: (1) calculating over the whole range of scales r the mean (μ) and the deviation (σ) of both C and D_C, and (2) breaking the set of scales into two distinct subsets $[r_{min}, \bar{r}]$ and $[\bar{r}, r_{max}]$, where $\bar{r}$ is the mean scale value, and calculating the corresponding means and deviations of D_C, in order to include local-scale information. Hence, the feature vector $CD = [CD_{1,\ldots,8}]$ is defined as

$$CD = \begin{matrix} [\mu(C), & \sigma(C) \\ \mu(D_C([r_{min}, r_{max}])) & \sigma(D_C([r_{min}, r_{max}])) \\ \mu(D_C([r_{min}, \bar{r}])) & \sigma(D_C([r_{min}, \bar{r}])) \\ \mu(D_C([\bar{r}, r_{max}])) & \sigma(D_C([\bar{r}, r_{max}]))] \end{matrix} \quad (7)$$

In order to explore the variation of the measurements either among different types of phonemes, or among phonemes that share similar phonetic characteristics, we measure the CD feature vector on a large set of embedded isolated phonemes from the TIMIT database (Garofolo et al., 1993), independently of the speaker sex or dialect; the amount of data used has on average order of magnitude of 2000 instances per phoneme. The measurements concern the univariate component densities so as to examine each component's relation to phonetic characteristics. The densities are shown in some cases in logarithmic scale for better visualization. The setup described holds for all succeeding density measurements.

In Fig. 2 we present indicative cases of histograms drawn from the CD feature vector such as the CD_5, referred to from now on as CD_{low}, that is the correlation dimension over the lower scales ($[r_{min}, \bar{r}]$), for selected phoneme types (see Fig. 2c and f). We observe that CD_{low} is higher for cases of strident fricative sounds (/s/,/z/), especially voiced ones, and lower for non-strident (/v/,/f/). Corresponding values for vowels seem to lie in between. Also, CD_{low} shows greater variance, and mainly lower values, for stops, with the voiced ordered higher than the unvoiced. Especially for the fricatives, as illustrated by Fig. 2f, it is

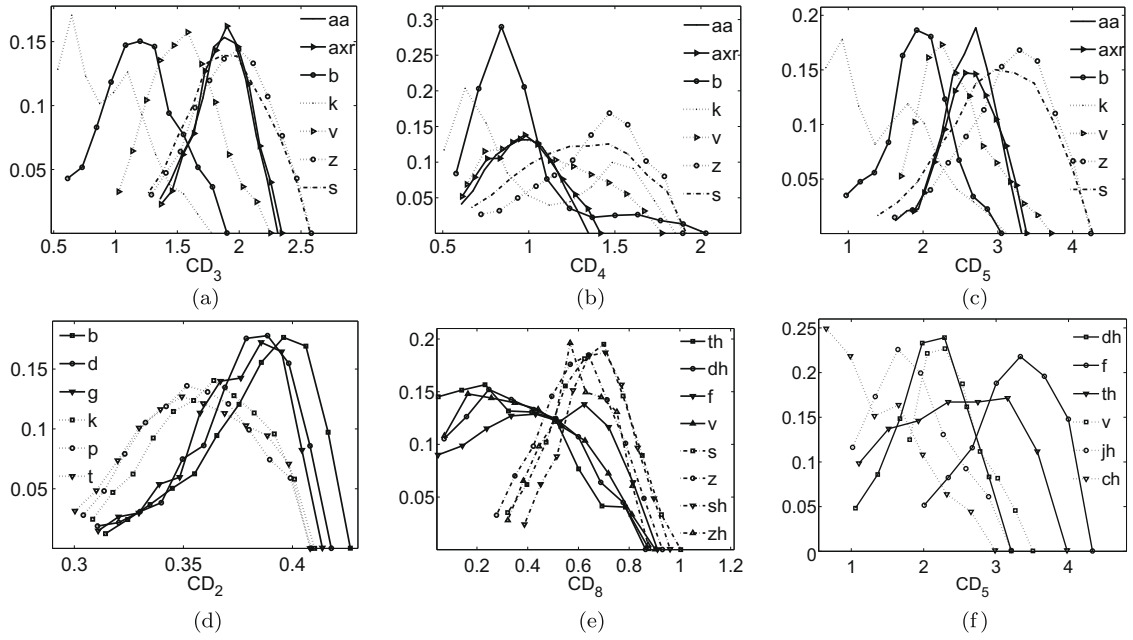


Fig. 2. Density of selected single-feature vector elements related to correlation dimension, indicative of phonemes from different classes: vowels, fricatives and stops. Top row: feature vector elements for selected phonemes from mixed classes are (a) $CD_3 = \mu(D_C)$, (b) $CD_4 = \sigma(D_C)$, (c) $CD_5 = \mu(D_C([r_{min}, \bar{r}]))$. Bottom row: cases of stops, affricatives and fricatives for (d) $CD_2 = \sigma(C)$, (e) $D_8 = \sigma(D_C([\bar{r}, r_{max}]))$, and (f) CD_5 feature vector elements.

observed that, among the non-strident ones, the voiced (e.g. /v/) demonstrate higher CD_{low} compared to the corresponding unvoiced ones (/f/) or alternatively to the affricatives (/jh/,/ch/). Furthermore on the fricatives, the deviation of the CD in higher scales (CD_8) is shown in Fig. 2e, assumes on average lower values for the non-strident fricatives (/th/,/dh/,/f/,/v/) versus their strident counterparts (/s/,/z/,/sh/,/zh/). Voiced stops (/b/,/d/,/g/) exhibit systematically different statistical characteristics than their unvoiced counterparts (/p/,/t/,/k/); this holds either for the CD_{low} or for the CD_2, as shown in Fig. 2d; the latter quantifies the spread of the correlation sum function over the range of scales. Other presented components include the average of the correlation dimension over all scales, and the corresponding deviation is shown in Fig. 2a and b, respectively.

3.2. Generalized dimensions

3.2.1. Background

The description of a phase-space via a single quantity, such as box-counting or correlation dimension, might not represent sufficiently a set since the underlying probability density may vary. Although fractal dimensions of the probabilistic type do take into account the variability of how often the system visits the different regions, they are a weighted average.

A method in the category of generalized dimensions of (Hentschel and Procaccia, 1983), which served as inspiration for the extension of the conducted measurements, is the generalized dimension function that defines an infinite class of dimensions, introduced in (Badii and Politi (1985)). This is accomplished by the computation of the moments of nearest neighbors' distances among randomly chosen points on the multidimensional set. Let $\delta(n)$ be the nearest neighbor distance among a reference point of the embedded set and the $n-1$ others, and $P(\delta,n)$ be the probability distribution of δ, then the moment of order γ of these distances is

$$\langle \delta^{\gamma} \rangle \equiv M_{\gamma}(n) = \int_0^{\infty} \delta^{\gamma} P(\delta,n) d\delta.$$

Since $\langle \delta^{\gamma} \rangle$ depends on n as $\sim n^{-\frac{\gamma}{D(\gamma)}}$ (Badii and Politi, 1985), the *dimension function* is defined as

$$D(\gamma) = -\lim_{n\to\infty} \frac{\gamma \log n}{\log M_{\gamma}(n)} \tag{8}$$

where γ is the parameter that suppresses or enhances the different distances of scale δ. Since for increasing γ the larger distances are more weighted and vice versa, $D(\gamma)$ is theoretically a monotonic non-decreasing function of γ. Among the infinite number of fractal dimensions with respect to the order parameter γ, one can find the Renyi class of dimensions D_q for $q \geqslant 0$. When $\gamma = (1-q)D_q$ the correspondence is realized as $D(\gamma) = D_q$. Geometrically the D_q's are the intersection of the $D(\gamma)$ graph with a set of lines with slope $\frac{1}{1-q}$. Thus, $D_{q=0}$ is the point that $\gamma = D(\gamma)$ and $D_{q=1}$ is the intersection with $\gamma = 0$. If $D(\gamma)$ is not varying with respect to γ, then the set is said to be homogeneous, with respect to the scales that are suppressed or amplified, possessing constant fractal dimension in the Renyi hierarchy of $D_q : D_0 = D_1 = \ldots = D_q, q \geqslant 0$, and vice versa.

The integral equation of $\langle \delta^{\gamma} \rangle$ can be rewritten as a sum for a discrete signal of finite length N:

$$M_{\gamma}(n) = \frac{1}{N} \sum_{i=1}^{N} \delta_i^{\gamma}(n) P(\delta_i, n) \tag{9}$$

where i is an index for the points of the data set. The probability density function $P(\delta,n)$ can be computed for an arbitrary scale δ_j as the difference of volume estimates based on the resolution of the successive scales (Hunt and Sullivan, 1986). Let $\{y(k) : k = 1,\ldots,M\}$ be a set of uniform random numbers of the same dimensionality as the data set X, and let us define the membership function $f_{\delta_j}(k) = 1$ if $\mathrm{dist}(y(k),X) \leqslant \delta_j$ and 0 otherwise where $\mathrm{dist}(y(k),X) = \inf_{x\in X}\|y(k) - x\|$. Then the volume estimate of a δ_j-cover of the set X is $A(\delta_j) \equiv \frac{1}{M}\sum_{k=1}^{M} f_{\delta_j}(k)$. Given the above, $P(\delta_j,N) \approx A(\delta_j) - A(\delta_{j+1})$ is an estimate of the probability that some point has a nearest neighbor at distance $\delta \in (\delta_{j+1},\delta_j]$. This probability $P(\delta_i,n)$ equalizes the corresponding nearest neighbor distances. The latter distances are computed among randomly sampled subsets of the original data. This procedure is repeated for the varying number of points that are included, in the considered subset giving rise to the n dependence, and for all the γ values, according to the above details, leading to the final moment M of order γ for varying number of points $M_{\gamma}(n)$.

3.2.2. Computation and intermediate measurements

In order to compute the dimension function we need to estimate the slope of $\log n^{\gamma}$ versus $\log M_{\gamma}(n)$ data. This is practically achieved by computing the mean slope of sequential estimations that result by a sliding window estimation within the range of $\log n^{\gamma}$ data. An indicative window utilized for slope estimation covers 7 points on the $\log M_{\gamma}(n)$ data.

Next we present intermediate measurements on the computation for a test data set, i.e. the Sinai toy system (see Fig. 3). The measurement of the generalized dimensions is conducted on two variants of data sampled from the uniform or the non-uniform Sinai system, respectively.[1] We show in the same plots a number of n^{γ} versus M_{γ} data points; in these we have subtracted the mean value of each one in order to make visualization feasible. Each group of points corresponds to a discrete γ value. Moreover, for each curve that corresponds to a different γ value there is superimposed the corresponding line-fit that shares the respective mean slope. This mean slope is considered as the average dimension with respect to each γ. Thus, by

[1] The Sinai 2D map is $(x_{n+1},y_{n+1}) = (x_n + y_n + g\cdot cos(2\pi y_n)\ mod\ 1,\ x_n + 2y_n)$, where $g = 0.3$ and $g = 0.02$ are the parameter values for the non-uniform and uniform cases, respectively.

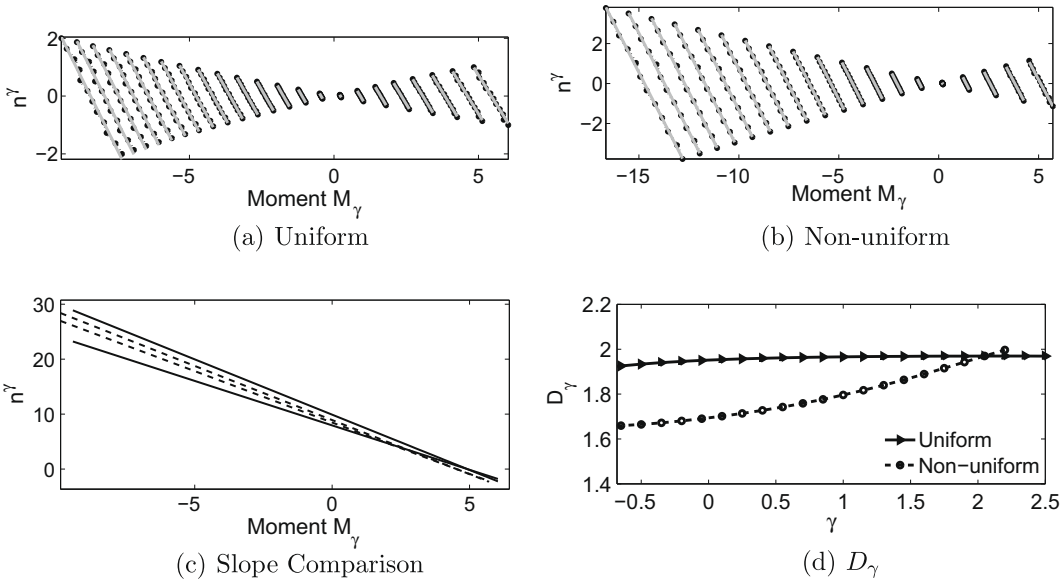


Fig. 3. Construction of the D_γ curve: slopes of the n^γ versus M_γ data points superimposed with the corresponding mean slope fits for the (a) uniform and (b) non-uniform Sinai system; and (c) explicit schematic comparison of the minimum and maximum slopes of (a) and (b) for the case of uniform (dashed line) versus the non-uniform case (solid line). (d) The resulting D_γ curves. (a) Uniform, (b) non-uniform, and (c) slope comparison D_γ.

computing the slope for each γ value we create the $D(\gamma)$ curve, henceforth referred to as D_γ for simplicity, shown in Fig. 3c. We also employ a schematic plot as shown in Fig. 3d in which we visualize via straight lines the minimum and maximum slopes for the two cases of system states, i.e. uniform versus non-uniform. Via this explicit comparison, the greater variation of the slope in the second case becomes apparent.

Following this procedure in Fig. 4b and d we show the corresponding measurements for two isolated phonemes after being embedded in the multidimensional phase-space. These correspond to two opposite cases: one in which the numerous dimensions for the subsequent γ values are constant, and to the other in which the dimensions vary with respect to the order parameter γ. In practice we use γ values in the range of $[-3, 3]$. However, the sparsity of the data points of the set might not allow the computation of the respective $D\gamma$ for all γ. This leads sometimes to possibly different domains in which the generalized dimension measurements are computed.

Among the practical issues, we should mention that the computational complexity of the fractal features is higher than the one of the MFCC. This is due mainly to the complexity of the embedding procedures including the computation of the embedding parameters; this complexity is increased by two orders of magnitude compared to the one of the cepstral features. However, computational issues can be further more radically accounted for by procedures discussed in (Kantz and Schreiber (1997)).

3.2.3. Comparison of measurements among fractal dimensions

We present next the geometrical correspondence among fractal dimensions from the Renyi hierarchy together with an indicative comparison among the fractal dimension-related measurements, i.e. the correlation dimension and the generalized dimensions. As mentioned in Section 3.2.1 this relation is given geometrically by the intersection of the graph of the D_γ function with a series of straight lines with slope $1/(1-q)$. Here, we show the generalized dimensions measurement for two cases of phonemes. In Fig. 5a along with the D_γ graph we have superimposed the lines that correspond to the cases of $q = 0, 1, 2$. It seems that the relatively constant case of the D_γ leads to almost equal D_q's: $D_0 \approx D_1 \approx D_2$. This is an example of the type of uniformity or homogeneity that we seek to detect since the dimension function is relatively constant. Besides, this measurement is close to the average correlation dimension (CD) that is superimposed in the same figure, as computed along the lines of Section 3.1. In this case the description of the set by a single quantity would be sufficient. Nevertheless, this information enriches our knowledge on the specific speech signal, since in the case that we would only have access to the single-valued correlation dimension we would not be aware of the set of values that the generalized dimensions render accessible. On the contrary, a different case is shown in Fig. 5b in which the characterization of the embedded phoneme is not sufficient by this single measurement, whereas the generalized dimensions vary with respect to the γ values. In this case the single CD measurement would not be sufficient.

3.3. GD features and characterization of speech signals

Based on the above raw measurements we construct a 9-element feature vector that is related to the generalized dimensions. This consists of: (1) the mean $\mu(D_\gamma)$ and the

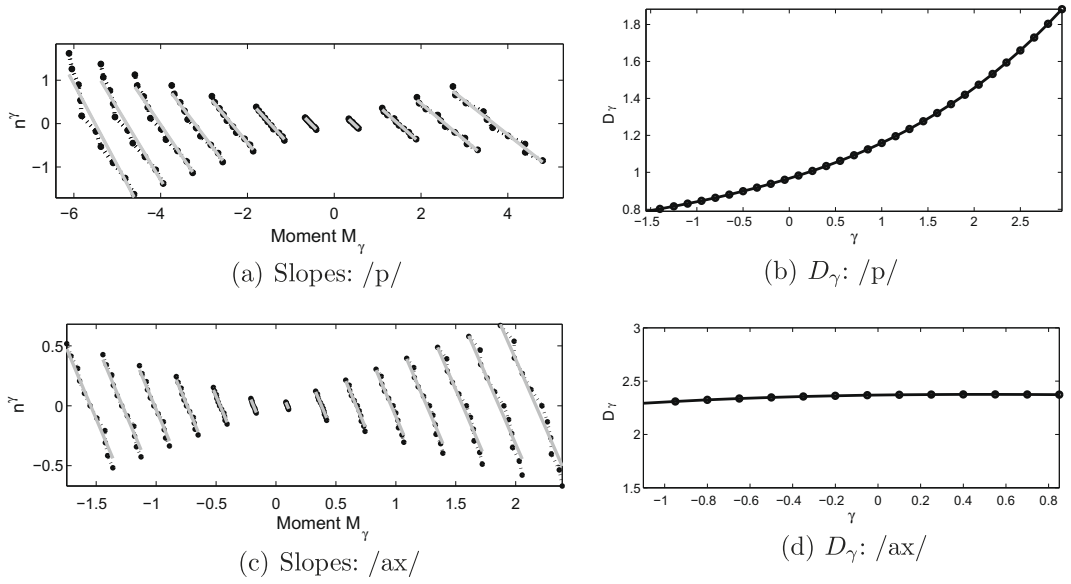


Fig. 4. Construction of the D_γ curve: (a and c) slopes of the n^γ versus M_γ data points superimposed with the corresponding mean slope fits; (b and d) the corresponding resulting mean slope points construct the D_γ curves for: top row, case of phoneme that has varying generalized dimensions (stop phoneme /p/); bottom row, the D_γ curve for the case of a phoneme for which it is relatively constant (vowel phoneme /ax/).

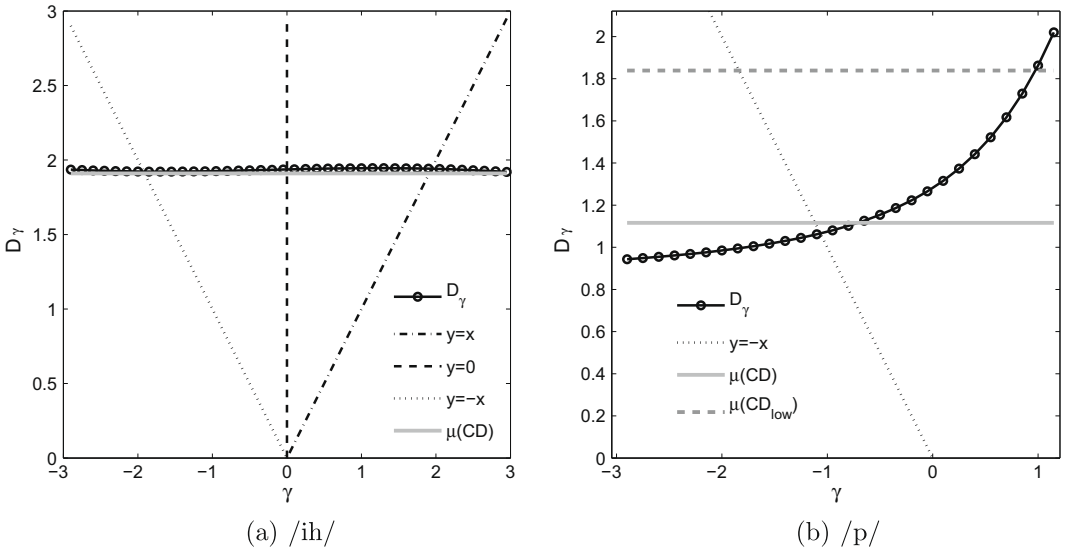


Fig. 5. Geometrical correspondence, for phonemes (a) /ih/ and (b) /p/, between the D_γ measurements and fractal dimensions from the Renyi hierarchy D_q: intersection points of the D_γ with lines of slopes equal to $\frac{1}{1-q}$ correspond to the D_q fractal dimension; $y=-x$, $y=0$, and $y=x$ for $q=2,1,0$, respectively. Indicative comparison with the average correlation dimension (CD) for the same phoneme type. (a) /ih/ (b) /p/.

standard deviation $\sigma(D_\gamma)$ of the dimension function, which include statistical information of the measurements; (2) the minimum, $\min(D_\gamma)$, and maximum, $\max(D_\gamma)$, values of the same function; (3) the parameters $[p_1, p_2, p_3]$ of a 2nd order polynomial fit $p_1 + p_2 \cdot \gamma + p_3 \cdot \gamma^2$ of the dimension function D_γ, which is also weighted by the corresponding estimation variances; these coefficients include more specific information on the location of the D_γ measurements and are thought of as the parametric decomposition of the D_γ into the specific basis; and (4) and finally, the boundaries $argmin_\gamma(D_\gamma)$ and $argmax_\gamma(D_\gamma)$ of the range of γ values for which the dimension function has been constructed. Hence, the generalized dimensions-related feature vector, referred to as GD, summarizes characteristics of the generalized dimensions and is defined as follows by its $GD_{1\ldots9}$ components:

$$GD = \begin{matrix} [\mu(D_\gamma), & \sigma(D_\gamma), & \min(D_\gamma), & \max(D_\gamma) \\ p_1, \quad p_2, & p_3, & \operatorname{argmin}_\gamma(D_\gamma), & \operatorname{argmax}_\gamma(D_\gamma)] \end{matrix} \tag{10}$$

Next we examine in detail how the distinct feature vector components are related to general phonetic characteristics, by examining their univariate densities.

3.3.1. Mean and variance feature components

Both the $\mu(D_\gamma)$ and $\sigma(D_\gamma)$ measurements are of interest: the mean dimension value is related to the values of the computed set of dimensions corresponding to an offset-like value over the generalized dimensions; in addition by focusing on the deviation, as if the above offset has been subtracted, low deviation suggests that the dimension function tends to be quite constant along the subsequent γ values and vice versa. By viewing the mean value $\mu(D_\gamma)$ (see 1st row of Fig. 6) for phoneme classes that share phonetic characteristics we observe the formation of statistical trends: the vowels have mean values in a specific range of values and their deviation (see Fig. 6 2nd row) is relatively low, compared to the one of the stops. Fricatives seem to share larger mean value again forming a discriminable statistical pattern for the cases of strong fricatives (/s/,/z/,/sh/, /zh/) as presented on the corresponding histogram. Taking a closer look, for example, at the stops we shall observe that among them the unvoiced ones versus the voiced ones follow two distinct trends, with the latter sharing broader distributed average values. Next, in Fig. 7, we see these statistical measurements superimposed for phoneme types that belong to different broad categories. Phonemes of the same broad class, sharing similar statistical characteristics, form densities that are consonant with each other; at the same time, these trends seem to be moderately distinguishable in some cases among the phonemes of the different type.

3.3.2. Lower and upper bound feature components

In the following we examine the minimum and maximum values of the generalized dimensions that represent a practical approximation to their lower and upper bounds. As Fig. 8a illustrates, the lower bound provides different

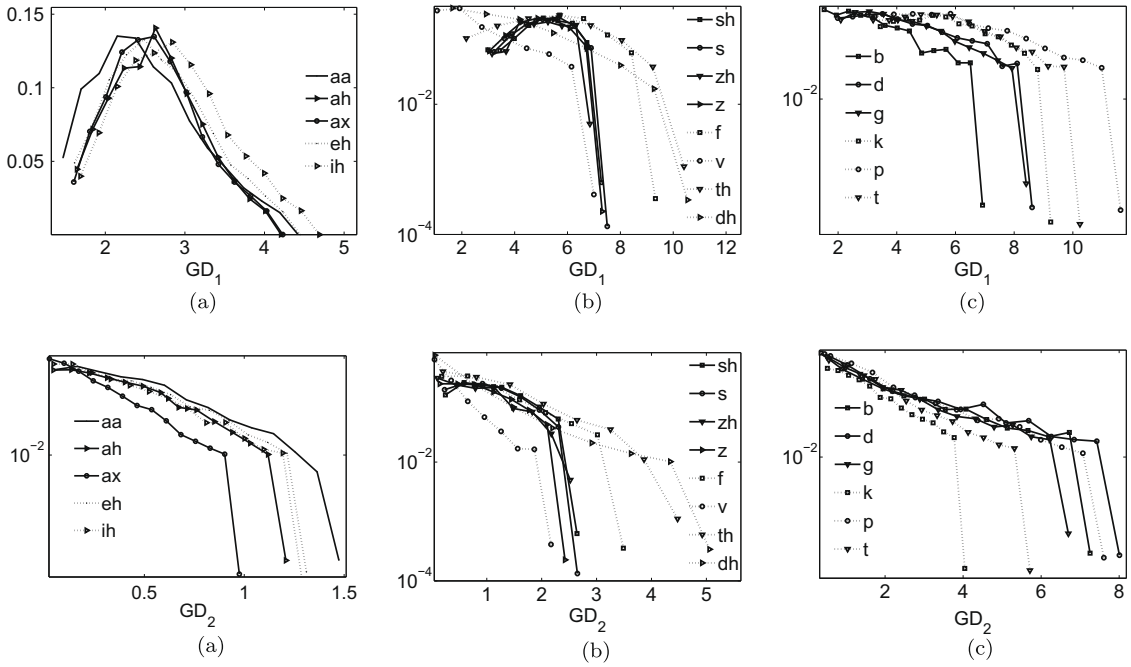


Fig. 6. Density of the mean GD_1 (top row) and the deviation GD_2 (2nd row) of the D_γ curves given the phoneme class; indicative of classes of (a) vowels, (b) fricatives, (c) stops.

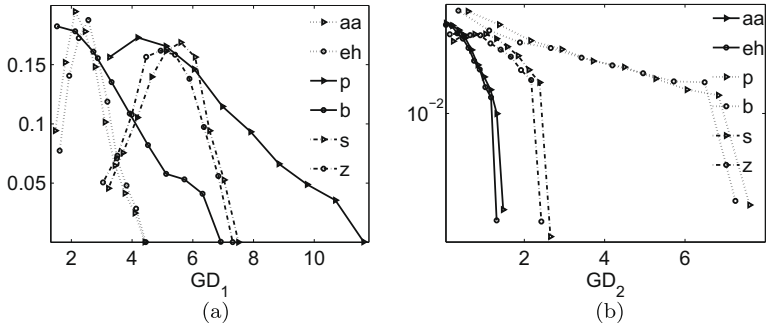


Fig. 7. Density of the (a) mean GD_1 and the (b) deviation GD_2 of the D_γ curves given the phoneme class; indicative of the different mixed classes of phonemes.

forms for the cases of voiced versus unvoiced stops. The same component, as pictured in Fig. 8b, differentiates slightly the densities of the front voiced fricatives (/v/, /dh/) from the corresponding unvoiced ones (/f/, /th/) or from the strong voiced ones (e.g. /z/) assigning on average lower values to the former. The vowels show lower values than most of the fricatives apart from the voiced fronts. The upper bound tends to lead to systematic forms in terms of statistical characteristics, demonstrating, as shown in Fig. 8c, greater values for the case of unvoiced fricatives, smaller for the voiced ones and even smaller for the vowels.

3.3.3. Polynomial decomposition feature components

The polynomial coefficient components of the GD feature vector are interpreted as a constant, a linear and a second order trend, all together approximating the D_γ function; moreover the constant term corresponds to an approximation of the information dimension D_I from the Renyi hierarchy, i.e. the value of the dimension function D_γ for $\gamma = 0$. We view next measurements on the three coefficients, denoted as $p_{1,2,3}$. The p_1 term shown in Fig. 9a seems to form statistical trends that differ either for the voiced non-strident fricatives (/v/, /th/) compared to either the unvoiced fricatives or to the voiced stridents (/z/, /zh/). Similar patterns are demonstrated in Fig. 9b among the vowels, the voiced stops, the unvoiced stops and the fricative unvoiced non-stridents or the voiced stridents. The values of the linear coefficient p_2, as pictured in Fig. 9c for the case of fricatives, show dependence on their type, for example, the front fricatives versus the strident fricative pairs. The p_3 term tends to lead to typical forms as shown in Fig. 9d demonstrating greater values for the unvoiced stops showing in addition higher variance. In the case of

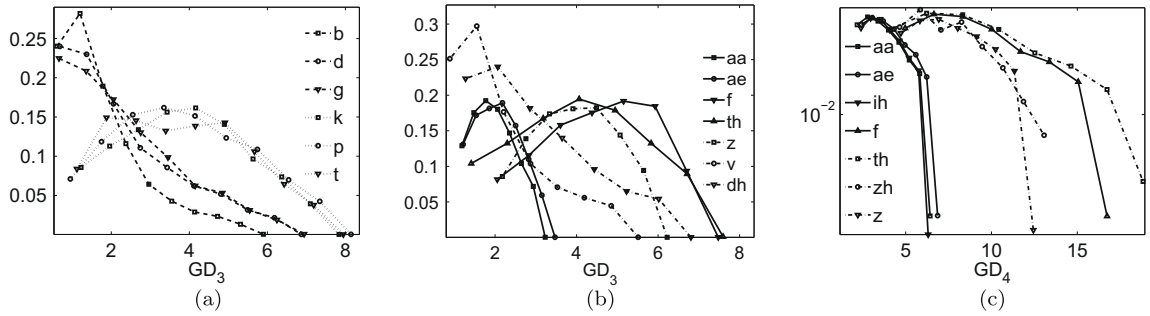


Fig. 8. Density of the lower bound GD_3 of the generalized dimension measurements for (a) stops and (b) mixed phoneme types. (c) Similarly the upper bound GD_4 in the case of mixed phoneme types.

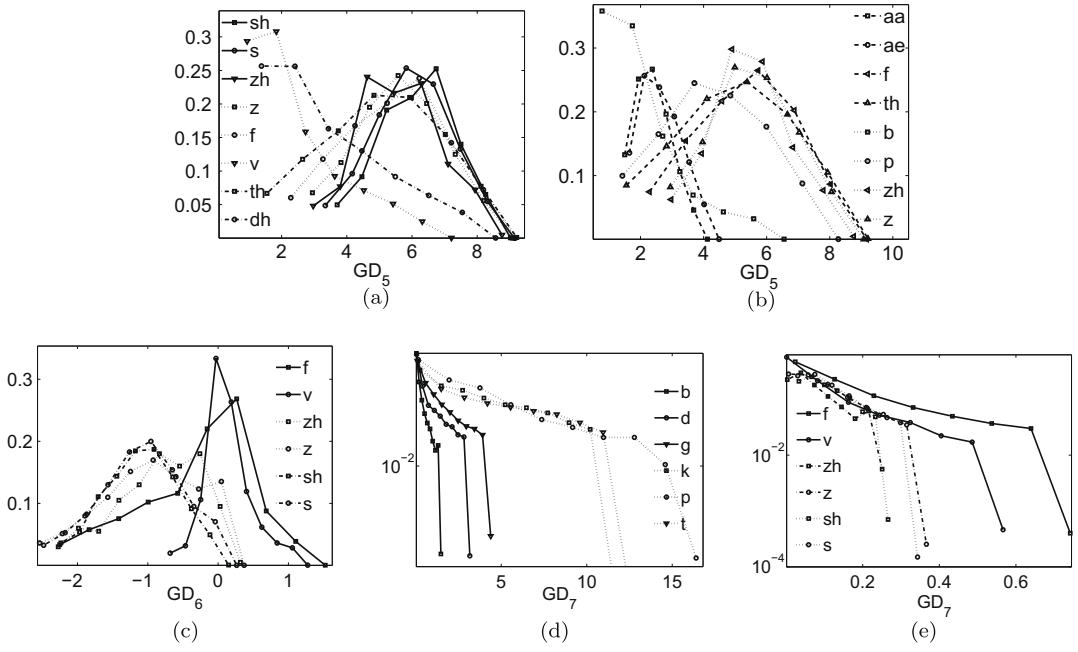


Fig. 9. Density of the components corresponding to the polynomial coefficients that decompose the generalized dimension function. Upper row: constant term DG_5 for (a) fricatives and mixed cases of phonemes. Bottom row: (a) linear term DG_6 for fricatives and (b and c) 2nd order term DG_7 for the same set of fricatives and stops.

fricatives, see Fig. 9e, the corresponding measurement gets lightly grouped, in terms of statistical characteristics, in pairs of related phonemes, such as the pair of front labials (/f/, /v/), the alveolar pair (/s/, /z/), and the palatal pair (/sh/, /zh/). Similar observations have been spotted among several types of phonemes, depending on the feature component utilized, as for instance, fricatives versus affricatives. However, an exhaustive enumeration of such properties is beyond our scope, since our goal is rather to expose indicative aspects on how the proposed measurements are related to the phonetic characteristics.

3.4. Comparison of features' statistical parameters

Finally, we present from a more macroscopic view, issues on the relation of the features' statistical parameters; at the same time we inspect in a more explicit way, the effect of phonetic characteristics on the features' statistical parameters. This is accomplished, in terms of the mean and variance of the distributions of the features, as follows: We question the normality or the log-normality on the univariate feature's phoneme distributions, employing hypothesis tests. For the cases that the null hypothesis is not rejected and that the realizations contained at least 100 entries, the mean and the variance of the corresponding phoneme's type distribution are estimated. In practice, this is the case for 90% of the distributions meeting the required constraint on the amount of data, whereas 76% among them were characterized as log-normal. The measurements are repeated across subsets formed by the eight speaker dialects of the TIMIT database, providing in this way multiple realization data.

On a second observation layer of the same results we superimpose in some indicative cases arrows that demonstrate roughly the effects of the general phonetic characteristics on the features' statistical parameters. These effects are observed due to a variation on a *single* characteristic while each time holding others constant. Such single characteristic variations refer for instance to the following: (1) The existence of voicing or not in the excitation (dashed lines); e.g. from /f/ to /v/ as shown in Fig. 10b corresponds to the case of varying the voicing while all other characteristics remain the same. (2) The manner of articulation (dotted lines), as for instance the variation among a stop, a fricative or a vowel; e.g. from /b/ to /v/ as shown in Fig. 10d corresponds to the transition from a stop to a fricative. (3) The place of articulation (full lines) such as the variation among a front, a central or a back; e.g. from /th/ to /f/ as shown in Fig. 10b that corresponds to the altering of the place from dental to labiodental. In this way one can see three types of "transitions" or "movements" in terms of the statistical parameters. Such an example is the existence of voicing that moves the parameters of the unvoiced stops or the unvoiced fricatives from right to the left in Fig. 10a; that is, showing lower mean. Similarly, variation on the place of articulation moves either the unvoiced stops or the front fricatives downwards, that is, altering their variance. Another case of movement due to the manner of articulation in the corresponding stop and fricative phonemes is shown either in Fig. 10c or d for the GD_3 or CD_5 component, respectively. In these cases we observe translation of the statistical measurements for two types of phonemes, from /d/ to /dh/ and from /b/ to /v/, i.e. altering the place of articulation while holding the other characteristics such as the voicing, or the manner of articulation the same. It seems that in many cases the variations of the statistical parameters of single-feature components form loose patterns due to the variation of phonetic characteristics. The numerical results advocate in favor of the previous observations. Moreover, they demonstrate some finer details concerning the statistics of the measurements. Indicative results are visualized in Fig. 10 by mean versus variance plots. In these for the shake of clarity the points that represent the multiple realizations – unless these are less than three data points – are represented by an ellipsis. Each ellipsis is centered on the center of mass of the data for each phoneme type, and its two axes are constructed according to the principal components of the underlying data. Each graph corresponds to the statistics of a single-feature component. Namely, the mean (Fig. 10a), the variance (b), the lower bound (c) of the generalized dimensions, and (d) the mean of the CD. It is shown by the conducted analysis that: (1) similar statistical trends demonstrated in the previous sections correspond to close points in the mean-variance scatter plots; (2) the positions of the phoneme parameters as visualized in these plots are related to their phonetic characteristics; and (3) the parameters for the different phoneme types are distinguishable in some cases with respect to the phoneme identity.

For instance, the statistics of the lower bound of the generalized dimensions, namely the GD_3 component, exhibit lower mean values for voiced versus their unvoiced phoneme cases. The vowels show less variance than consonants. In the case of fricatives the place of articulation causes similar statistics on the variance of the GD_1 component; assigning either in fricatives or in stops from higher to lower variance on fronts and backs, respectively. In the same graph unvoiced stops tend to concentrate on the right upper corner, voiced ones left, front coming first, followed by the central and back ones to its right.

4. Correlation between fractal and cepstral features

With the presented perspective, which employs the fractal features, we attempt to measure information, which cepstral-originated features might not represent in specific cases. Towards this direction, next we shed some light on issues concerning the linear correlation between the fractal and MFCC features. In the following, that also indicate a new approach perspective, firstly, we discuss qualitatively the correlation between the features of the different type with respect to the phoneme classes; secondly, we reconstruct in the same context the spectral content that

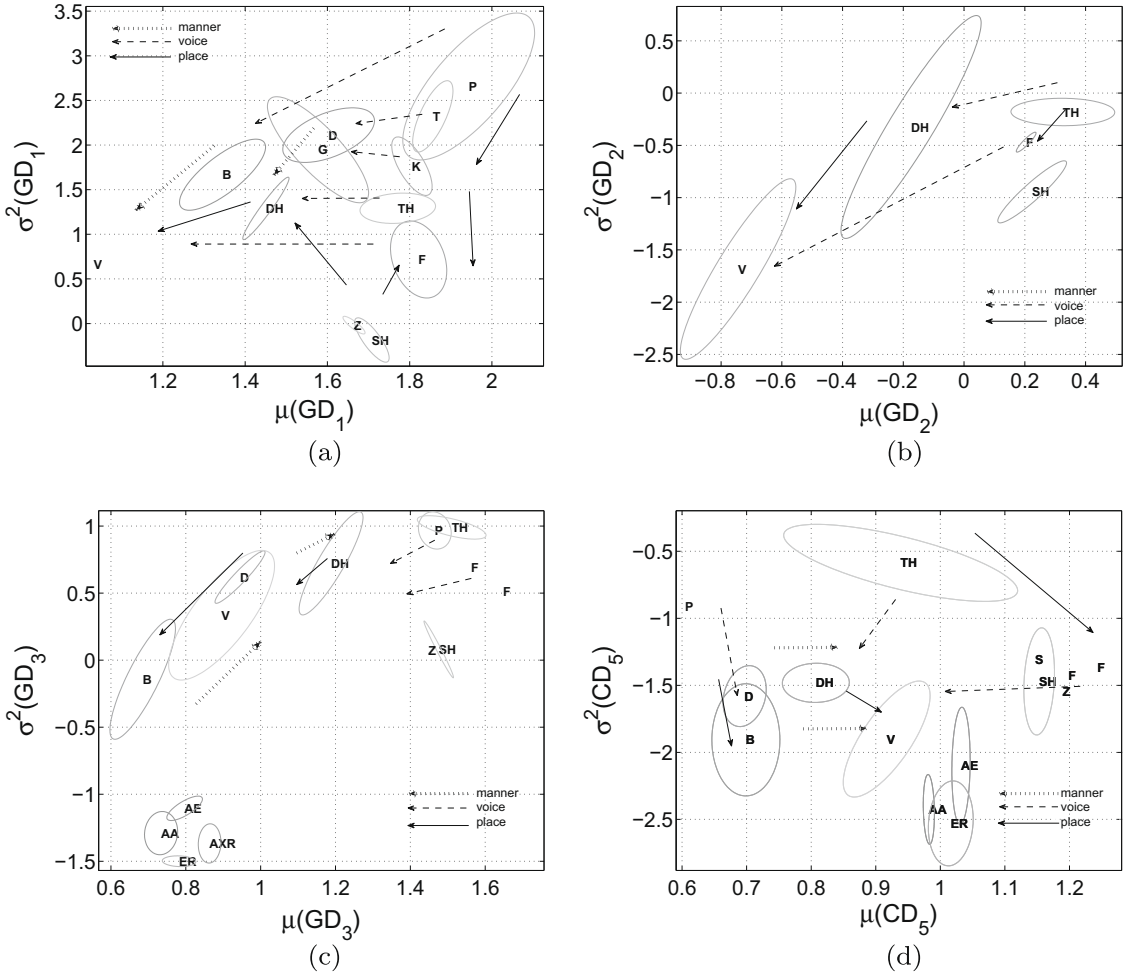


Fig. 10. Mean versus variance scatter plots of single component feature statistics. After multiple realizations that correspond to the different speaker dialects of the TIMIT database, an ellipsis is fitted on the underlying data points that each represent the parameters of each dialect's distribution for the corresponding phoneme type. Line arrows illustrate the effect on the statistical parameters of the represented feature component when varying a single phonetic characteristic each time such as: The existence of voicing or not in the excitation (dashed lines), e.g. from /f/ to /v/ as in (b). The manner of articulation (dotted lines), e.g. from /b/ to /v/ as in (d). The place of articulation (full lines), e.g. from /th/ to /f/ as in (b). Components shown include (a) the GD mean: GD_1, (b) the GD variance: GD_2 (c) the GD lower bound: GD_3 and (d) the CD mean in lower scales: CD_5. Refer also to Section 3.4.

corresponds to the fractal features regarding the most and least correlated components with the MFCC.

4.1. Correlation with respect to the phonemes

Towards the exploration of the linear correlation between the fractal and MFCC features we employ canonical correlation analysis (CCA) (Anderson, 2003); in this way we create two bases, one for each feature set, i.e. the MFCC and the CD fractal-related feature vectors. These bases are developed so that their eigenvectors are ordered from the most correlated ones, among the two feature sets, to the least correlated ones. Next, we compute the sorted eigenvalues for the two feature vectors with respect to the different phoneme types separately for each speaker.

Fig. 11a and b visualizes the measurements, showing the correlation coefficient among the two feature sets while this varies with respect to the phoneme type. The phoneme type is represented in *sorted* order, based on average values, from the least correlated to the most correlated one. For example, certain phoneme types hold larger on average coefficients but show lower values in the less correlated components, that fall sharply in the following components; others may retain their modest correlation coefficient across more components. Such a case is formed between /p/ and /ih/ in Fig. 11. In general it seems that across speakers the unvoiced fricative and stop phonemes are ordered lower in terms of correlation, i.e. to the left of the x-axis as shown in the graphs, than vowels. Among the latter, the back vowels of the /i/-class are ordered once again, lower than the others. Similar patterns, i.e. more correlated components for some phoneme types, for example /aa/-like vowels and less correlated components in others such as some fricatives or unvoiced stops, are observed across groups of different speakers.

Given the sorting according to the average correlation coefficient of the phonemes we proceed by computing the density with respect to the phoneme type of the phonemes

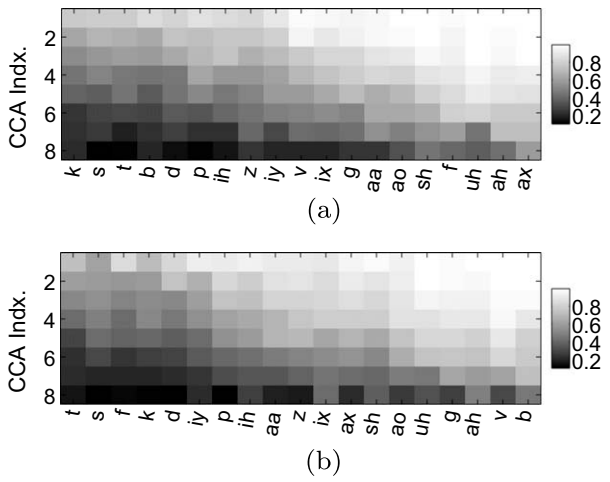


Fig. 11. Canonical correlation coefficients measurements between the feature vector components of the MFCC and the fractal features with respect to the phoneme type (x-axis). The coefficients' values are in grayscale shown in the side-bar. The coefficient index of the y-axis has the minimum rank among the feature vectors. Results are shown for two different speakers; speaker identities are (a) mdns0 and (b) mkls0. Phoneme labels in the x-axis are sorted with respect to the average correlation index per phoneme; data are from the TIMIT database.

that are ranked as 1st or 2nd lowest according to their correlation across all speakers. In this experiment all speakers of the TIMIT database are taken into account. The results, visualized in Fig. 12, indicate that the correlation between the two feature sets, i.e. fractals and MFCC, varies with respect to the phoneme class; additionally this variation seems to follow certain phoneme-wise patterns. Unvoiced stops and fricatives are ranked as the lowest correlated, followed by the back vowels. Moreover,from the examined viewpoint the above implicitly supports that the proposed features contribute non-correlated information that depends on the phoneme type. Next, we continue on an effort to view aspects of this correlation information by demonstrating maximal or minimal cases in terms of the spectral content of speech.

4.2. Fractal features' spectral content

Thereafter, we explore another aspect concerning the correlated parts of information between the fractal and the MFCC features; this focuses on the spectral content of their most and least correlated components. We employ once again the CCA, however, in the utterance level and independent to the phoneme type. At first, we reconstruct the truncated spectra of an utterance, denoted by F_{trunc}, by utilizing the corresponding MFCC features as shown in Fig. 13a. Next, given the ordering among the most and least correlated components of the learned canonical bases, we keep only the most correlated one; then, we reconstruct the corresponding MFCC features by utilizing the CCA learned basis of the *fractal features* and mapping them back to the MFCC vector space. At this point we reconstruct their spectra F_{most} in order to be comparable with the original spectra (see Fig. 13b). This comparison shows that this mostly correlated component owes its correlation mostly to the lower frequency content retaining time relative information; in frequencies greater than roughly $\frac{1}{4}$ times the sampling frequency its spectral content is flattened across time in different frequency bands; an effect of the scalewise processing that the fractal features undergo. Next, we repeat the above procedure, by keeping

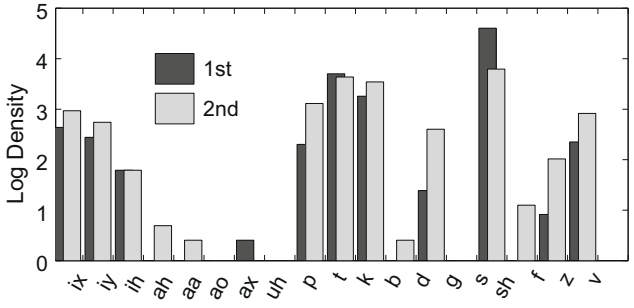


Fig. 12. Histograms in logarithmic scale for (a) 1st and (b) 2nd ranked phonemes across all speakers; ranking is defined in terms of lowest correlation among the included phonemes.

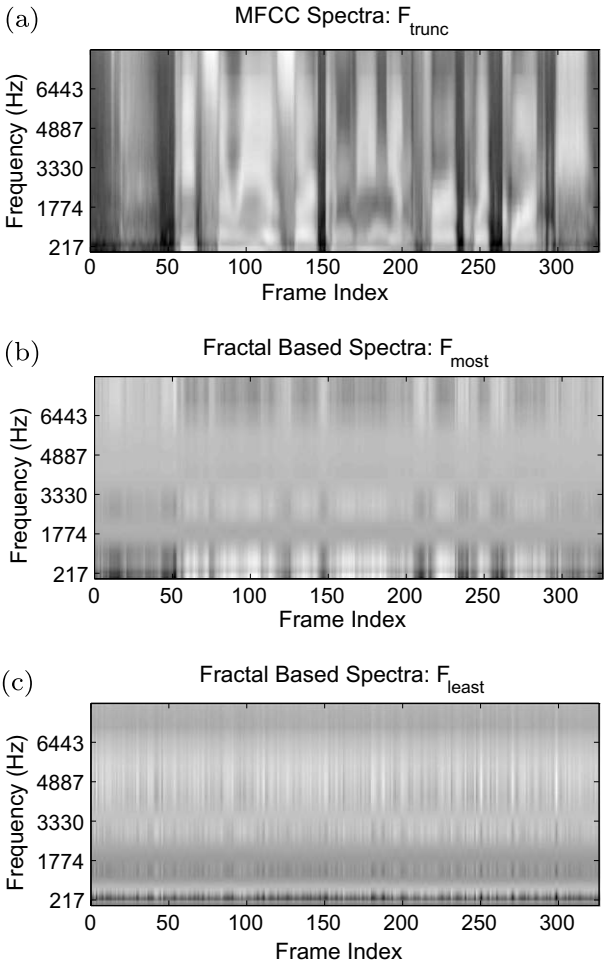


Fig. 13. (a) Original spectra that correspond to the 13-element MFCC feature vector, reconstructed spectra of the (b) *most* and (c) *least* correlated components between the MFCC and fractal features.

this time only the least correlated component of the CCA learned fractal features' basis. This is used to transform the fractal features to the MFCC vector space and on their turn to be used for the spectra reconstruction of F_{least} presented in Fig. 13c. This experiment shows that the least correlated component's spectral content is shown to lose the time–frequency information compared to the original spectra. This observation highlights that the least correlated information between the fractal and MFCC features does not show any structure concerning its spectral content.

Given the variation with respect to the phoneme type of the correlated components between the fractal and the MFCC features in Section 4.1, we have shown maximal and minimal correlation cases on the information that the different features share in terms of their spectral content. At the same time, the aspects examined in the above qualitative analysis illustrate properties of the features employed as far as the relation to well-known quantities such as the spectrum are considered.

5. Fractal features in phoneme classification

A first set of experiments is conducted on sets of single phonemes. This allows us to inspect the phoneme confusability among the different classes; however, the phonemes included are restricted to the union of the subsets that are part of the specific task. Next, we restrict the set-up by considering phoneme classification in broad classes; this is realized by merging phonetically proximate classes.

On the proposed analysis we require each embedded and further processed signal to be a complete phoneme. This implies that each phoneme shall correspond to a single-feature vector. On the other hand MFCC features are based on short-time processing so as to account principally for non-stationarity, suggesting frame-wise features. To account for this heterogeneity in terms of comparison, apart from the frame-wise MFCC baseline that exploits dynamical information, we also compare in some cases the results of the fractal features with a second variant of MFCC-based baseline. The latter utilizes an average with respect to the cepstral coefficients so as to map all frames into one.

The speech corpus utilized is the TIMIT database (Garofolo et al., 1993), which is accompanied by hand-labeled phoneme-level transcriptions. Each signal processed is an *isolated* phoneme. The training and testing sets have been employed as are defined in the original speaker independent setup. The classification experiments make use of the partitioning of phonemes into broad categories. These classes are vowels (Vo), fricatives (Fr), stops (St), nasals (Na), liquids (Li), voiced (Voi), unvoiced (Un), fronts (Fro), centrals (Ce) and backs (Ba); the specific phonemes that each category consists of are listed in Table 1.

5.1. Single phoneme classification and confusability

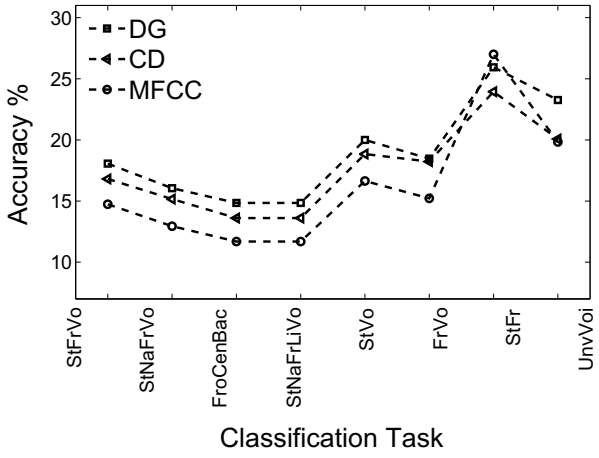
At first we examine the classification efficacy of each fractal feature among single phonemes that are contained in a specific subset. The subsets used are unions of the sets defined in Table 1. The acoustic modeling has been realized using the HTK (Young et al., 2002) with 1-state Hidden Markov Models (HMMs) for the fractal and the MFCC features that contain each a *single* feature vector per phoneme. In detail, we next show in Fig. 14a the classification accuracies for each one of the CD and the GD feature vectors across the various scenarios that are enumerated along the x-axis. The accuracies for each classification task among the single phoneme classes range from 12% to 28%, depending on the set of phonemes considered. The single phoneme classification experiment allows us to observe the confusability within the different phoneme classes across various scenarios. This is visualized in the confusion matrix shown in Fig. 14b that corresponds to the classification task among all phonemes contained in either one of the classes of stop or vowels. We observe that the intra-class confusability for either the vowels or the stops is higher than for other cases. Another confusable intra-group is formed among the unvoiced stops; similar results have been observed in other scenarios too. Given these observations we proceed and take the union of phoneme sets into broad classes that share phonetic characteristics.

5.2. Broad class phoneme classification

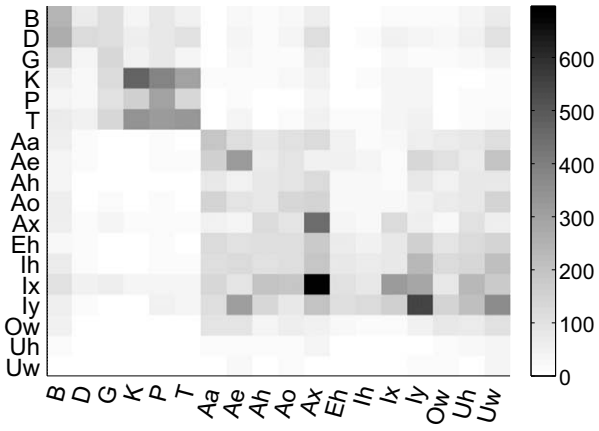
In a first set of experiments we focus on each single-feature vector component of the fractal features. The acoustic modeling has been realized with 1-state HMM for the fractal features. We next show in detail in Fig. 15a the classification

Table 1
Partitioning of phonemes into broad classes.

Type	Abrv.	Phonemes																	
Vowel	Vo	aa	ae	ah	ao	ax	eh	ih	ix	iy	ow	uh	uw						
Fricative	Fr	ch	dh	f	jh	s	sh	th	v	z	zh								
Stop	St	b	d	g	k	p	t												
Nasal	Na	em	en	m	n	ng													
Liquid	Li	el	hh	l	r	w	y												
Front	Fro	ae	b	eh	em	f	ih	ix	iy	m	p	v	w						
Central	Ce	ah	ao	axr	d	dh	el	en	er	l	n	r	s	t	th	z	zh		
Back	Ba	aa	ax	ch	g	hh	jh	k	ng	ow	sh	uh	uw	y					
Voiced	Voi	b	d	dh	el	em	en	g	jh	l	m	n	ng	r	v	z	zh	w	y
Unvoiced	Uv	ch	f	hh	k	p	s	sh	t	th									

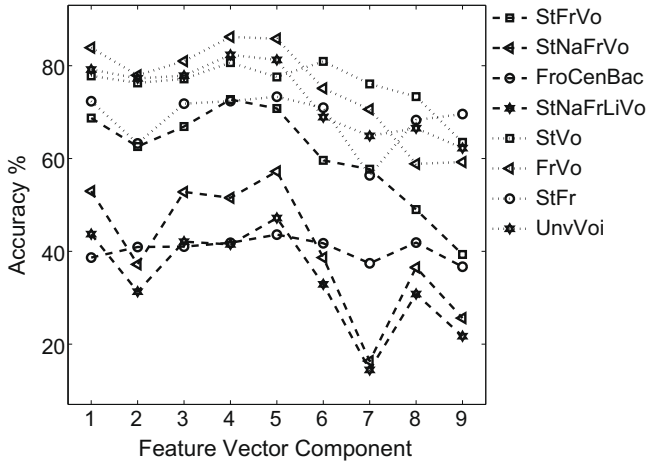


(a) Single Phoneme Classification

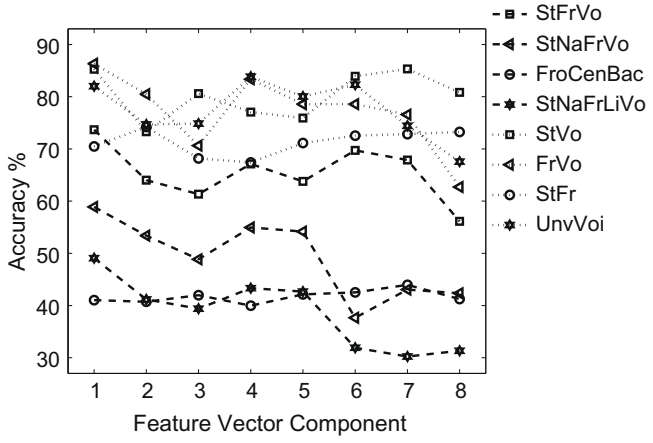


(b) Confusion Matrix

Fig. 14. (a) Classification accuracy in single phonemes among the phonemes that are contained in the classes of each classification task, (x-axis) for each CD, GD and MFCC feature vector; baseline MFCC features are averaged and mapped in a single frame per phoneme. (b) Confusion matrix, for the feature component CD_5, of the 5th classification task among the scenarios included in (a), i.e. stop versus vowels on single phonemes. We observe the higher confusability among phonemes sharing similar characteristics such as stops, a subset of vowels or unvoiced stops.



(a) GD



(b) CD

Fig. 15. Classification accuracy in broad phoneme classes for each single component (x-axis) of the (a) GD and the (b) CD feature vector; the classification tasks appear in the legend. (a) GD, and (b) CD.

accuracies for each one of the GD components, and in Fig. 15b we show the corresponding accuracies for the isolated CD feature vector elements. It seems that among the fractal feature components some are more efficient in certain classification tasks than others. For instance, the constant term of the polynomial decomposition (see Eq. 10), that is the GD_5 component, performs better than the linear term of the decomposition, that is the GD_6 component, in the fricatives versus vowels scenario compared to the stops versus vowels scenario and vice versa. Another case that demonstrates different performance between a scenario that is based on the discrimination given the manner of articulation and a scenario that represents the discrimination depending on the existence of voicing or not is the case of the CD_6 and CD_8 feature elements (see Eq. 7): the former component performs better in the case of the second scenario (voicing) and vice versa. The same holds for the GD_7, GD_9 pair.

In the second set of experiments we have explored the classification efficacy of the whole feature vectors. Moreover, we have also employed for comparison the more advanced baseline. The acoustic modeling in this case has been realized with 1-state HMM for the fractal features that contains a single-feature vector for each phoneme, and 3-state HMM for the MFCC. The latter contain multiple frame-wise feature vectors per phoneme and are augmented by derivative and acceleration coefficients, taking advantage of the phoneme dynamics. The various classification scenarios highlight from a different viewpoint the characteristics of the features relative to the phoneme classes they are called to represent.

The classification scores for the 8 experiments shown in Table 2 indicate the capability of the proposed feature sets to classify phonemes into broad classes; some cases such as Voiced versus Unvoiced phonemes perform better than Front versus Central versus Back phoneme classes. Whilst

Table 2
Classification scores (%) for broad phoneme classes[a] using either the MFCC baseline features or plainly the fractal features. MFCC features are computed framewise for each phoneme. Fractal features are correlation dimension (CD), Generalized Dimensions (GDs). CD + GDs label stands for the concatenation of the corresponding feature sets.

	St/Fr/Vo	St/Na/Fr/Vo	Fro/Ce/Ba	St/Na/Fr/Li/Vo
MFCC	88.83	84.87	61.18	75.05
CD	81.62	68.87	40.33	55.84
GD	84.65	69.89	43.04	58.02
CD + GD	87.29	75.35	42.98	61.54
	St/Vo	Fr/Vo	Un/Vo	Un/Voi
MFCC	93.61	93.41	93.09	83.29
CD	94.49	86.83	91.85	83.92
GD	95.20	89.97	94.79	86.62
CD + GD	96.62	92.93	97.07	89.05

[a] Classes are vowel (Vo), fricative (Fr), stop (St), nasal (Na), liquid (Li), voiced (Voi), unvoiced (Un), front (Fro), central (Ce) and back (Ba).

the fractal features alone contain 8 and 9 components *per phoneme*, for the correlation dimension (CD) and generalized dimension (GD) features, respectively, they occasionally yield comparable accuracies to the MFCC feature vector containing 39 coefficients *per frame*. Apart from the different, to a certain degree, information that the fractal features carry, this fact could be also considered as an indication of a more economic representation of the broad phoneme types. Another issue to notice is that the generalized dimensions-related feature set performs better than the correlation dimension feature set. The average performance over the presented classification scenarios of the GD features is 77.8%, compared to 75.5% of the CD features, and 84.2% of the MFCC. Finally, when the CD and GD features are combined by simple concatenation in a single-feature vector they perform modestly better than either cases in which they have been employed on their own, showing average performance of 80.4%.

6. Conclusions

In this paper we present the application of speech signal processing methods inspired by dynamical systems and fractal theory for the analysis and characterization of speech sounds. The steps taken consist of the embedding procedure that constructs a multidimensional space, followed by measurements related to the correlation dimension and generalized dimensions for the practical cases of speech signals. Then, we utilize these measurements to extract simple feature vectors. The analysis of the features in terms of their statistical trends has shown them to form statistical patterns depending on their general phonetic characteristics. For instance distinct feature vector elements obtain on average values that are subject to characteristics such as the voicing, the manner and the place of articulation. Moreover the variation of the statistical parameters of the features seems to follow loose-formed patterns when we alter a single phonetic characteristic (e.g. place of articulation). These patterns seem to be similar in different types of phonemes, e.g. fricatives or stops. Next, we employ a variety of classification experiments, primarily among broad phoneme types. Both the intermediate statistical measurements together with the qualitative analysis and quantitatively the classification experiments indicate that the information carried by the extracted features characterizes to a certain extent the different speech sound classes. The quantitative results are comparable occasionally with the baseline features' classification results; at the same time the features consist of much smaller number of feature components. Another issue addressed, which has not been considered up to now, is the varying correlation with respect to the phoneme type between the fractal features and the MFCC. This is explored by means of canonical correlation analysis, and shows lower correlation coefficients for unvoiced stops and fricatives, or backs in the case of vowels compared to other types of phonemes. Continuing the above, in this light, we also examined this varying correlation information in terms of the spectra of the most and least correlated components. This direction concerns an aspect of the fractal features' spectral content and lets us observe a concentration of the least correlated information in bands that span the higher frequencies lacking any time-frequency structure, whilst the most correlated components mainly contain lower spectral content also characterized by time-related resolution.

The specific fractal features cannot be compared with the baseline MFCC features in terms of classification experiments, as far as the resulting accuracy is concerned. This raises a number of issues that someone would consider to look into. Among the most important issues resides the subject of fusion between the feature that carries the first order information as considered in this work, i.e. the MFCC, and the nonlinear features, which are considered to carry second order information. On previous works we have considered simple fusion approaches (Maragos and Potamianos, 1999; Pitsikalis and Maragos, 2006). Interesting research directions involve the exploitation of concepts of adaptive fusion by uncertainty compensation (Papandreou et al., 2009), by modeling multiple sources of uncertainty, such as measurement or model uncertainty; such an approach has been explored for the case of audio and visual streams for audio-visual classification. Towards this direction, it seems that it would also be worth exploring aspects of the correlation among the multiple features: we think of expanding the ideas presented in Section 4 of this paper, by use of the canonical correlation analysis so as to take advantage of the varying correlation among the different models. From another viewpoint it would be interesting to consider the problem of fusion at the front-end level, by incorporating the multiple types of information in a common algorithm: spectral information together with information related to complexity quantification. As far as the statistical modeling for fusion of the multiple feature cues is concerned, state-synchronous modeling does not fit on

the *specific* phoneme-level approach concerning the fractal features. In contrast, models such as the parallel-HMM, or other generalizations (Potamianos et al., 2004), could be more appropriate. Finally, an interesting track for further research includes the investigation of the relation of fractal measurements with concepts that are more related to the physics of speech production. Towards this direction, one could explore the association of the proposed methods with concepts from the area of articulatory characteristics of speech production (Deng et al., 1997; Livescu et al., 2003).

Acknowledgements

This work was supported in part by the FP6 European research programs HIWIRE and the Network of Excellence MUSCLE and by the 'Protagoras' NTUA research program.

References

Abarbanel, H.D.I., 1996. Analysis of Observed Chaotic Data. Springer-Verlag, New York, Berlin, Heidelberg.

Adeyemi, O., Boudreaux-Bartels, F.G., 1997. Improved accuracy in the singularity spectrum of multifractal chaotic time series. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, Germany, pp. 2377–2380.

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Badii, R., Politi, A., 1985. Statistical description of chaotic attractors: the dimension function. J. Statist. Phys. 40 (5–6), 725–750.

Banbrook, M., McLaughlin, S., Mann, I., 1999. Speech characterization and synthesis by nonlinear methods. IEEE Trans. Speech Audio Process. 7 (1), 1–17.

Benzi, R., Paladin, G., Parisi, G., Vulpiani, A., 1984. On the multifractal nature of fully developed turbulence and chaotic systems. J. Phys. A 17, 3521–3531.

Bernhard, H.-P., Kubin, G., 1991. Speech production and chaos. In: XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France, pp. 19–24.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Process. 28 (4), 357–366.

Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. Speech Commun. 22, 93–111.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium, Philadelphia.

Grassberger, P., Procaccia, I., 1983. Measuring the strangeness of strange attractors. Physica D 9 (1–2), 189–208.

Greenwood, G.W., 1997. Characterization of attractors in speech signals. BioSystems 44, 161–165.

Hentschel, H.G.E., Procaccia, I., 1983. Fractal nature of turbulence as manifested in turbulent diffusion strange attractors. Phys. Rev. A 27, 1266–1269.

Hentschel, H.G.E., Procaccia, I., 1983. The infinite number of generalized dimensions of fractals and strange attractors. Physica D 8 (3), 435–444.

Herzel, H., Berry, D., Titze, I., Saleh, M., 1993. Analysis of vocal disorders with methods from nonlinear dynamics. NCVS Status Progress Rep. 4, 177–193.

Hirschberg, A., 1992. Some fluid dynamic aspects of speech. Bull. Commun. Parlée 2, 7–30.

Howe, M.S., McGowan, R.S., 2005. Aeroacoustics of [s]. Proc. Roy. Soc. A 461, 1005–1028.

Hunt, F., Sullivan, F., 1986. Efficient algorithms for computing fractal dimensions. In: Mayer-Kress, G. (Ed.), Dimensions and Entropies in Chaotic Systems. Springer-Verlag, Berlin.

Johnson, M.T., Povinelli, R.J., Lindgren, A.C., Ye, J., Liu, X., Indrebo, K.M., 2005. Time-domain isolated phoneme classification using reconstructed phase spaces. IEEE Trans. Speech Audio Process. 13, 458–466.

Kaiser, J.F., 1983. Some observations on vocal tract operation from a fluid flow point of view. In: Titze, I.R., Scherer, R.C. (Eds.), Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control, Denver Center for Performing Arts, Denver, CO, pp. 358–386.

Kantz, H., Schreiber, T., 1997. Nonlinear Time Series Analysis. Cambridge University Press, Cambridge, UK.

Kokkinos, I., Maragos, P., 2005. Nonlinear speech analysis using models for chaotic systems. IEEE Trans. Acoust., Speech, Signal Process. 13 (6), 1098–1109.

Kubin, G., 1996. Synthesis and coding of continuous speech with the nonlinear oscillator model. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-96, Vol. 1, Atlanta, USA, p. 267.

Kumar, A., Mullick, S.K., 1996. Nonlinear dynamical analysis of speech. J. Acoust. Soc. Am. 100 (1), 615–629.

Livescu, K., Glass, J., Bilmes, J., 2003. Hidden feature models for speech recognition using dynamic bayesian networks. In: 8th European Conference on Speech Communication and Technology (Eurospeech), pp. 2529–2532.

Mandelbrot, B., 1982. The Fractal Geometry of Nature. Freeman, NY.

Maragos, P., 1991. Fractal aspects of speech signals: dimension and interpolation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-91, pp. 417–420.

Maragos, P., Potamianos, A., 1999. Fractal dimensions of speech sounds: computation and application to automatic speech recognition. J. Acoust. Soc. Am. 105 (3), 1925–1932.

Maragos, P., Kaiser, J.F., Quatieri, T.F., 1993. Energy separation in signal modulations with application to speech analysis. IEEE Trans. Signal Process. 41 (10), 3024–3051.

Meneveau, C., Sreenivasan, K.R., 1991. The multifractal nature of turbulent energy dissipation. J. Fluid Mech. 224, 429–484.

Narayanan, S., Alwan, A., 1995. A nonlinear dynamical systems analysis of fricative consonants. J. Acoust. Soc. Am. 97 (4), 2511–2524.

Packard, N., Crutchfield, J., Farmer, D.S.R., 1980. Geometry from a time series. Phys. Rev. Lett. 45, 712–716.

Papandreou, G., Katsamanis, A., Pitsikalis, V., Maragos, P., 2009. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Trans. Audio, Speech Language Process. 17 (3), 423–435.

Peitgen, H.-O., Jürgens, H., Saupe, D., 1992. Chaos and Fractals: New Frontiers of Science. Springer-Verlag, Berlin, Heidelberg.

Pitsikalis, V., Maragos, P., 2002. Speech analysis and feature extraction using chaotic models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-02, Orlando, USA, pp. 533–536.

Pitsikalis, V., Maragos, P., 2006. Filtered dynamics and fractal dimensions for noisy speech recognition. IEEE Signal Proc. Lett. 13 (11), 711–714.

Pitsikalis, V., Kokkinos, I., Maragos, P., 2003. Nonlinear analysis of speech signals: generalized dimensions and Lyapunov exponents. In: Proceedings of the European Conference on Speech Communication and Technology, Eurospeech-03, Geneva, Switzerland, pp. 817–820.

Potamianos, G., Neti, C., Luettin, J., Matthews, I., 2004. Audio-visual automatic speech recognition: an overview. In: Bailly, G., Vatikiotis-Bateson, E., Perrier, P. (Eds.), Issues in Visual and Audio-Visual Speech Processing. MIT Press (Chapter 10).

Quatieri, T.F., Hofstetter, E.M., 1990. Short-time signal representation by nonlinear difference equations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'90, Albuquerque, Vol. 3, pp. 1551–1554.

Sauer, T., Yorke, J.A., Casdagli, M., 1991. Embedology. J. Statist. Phys. 65 (3–4), 579–616.

Takens, F., 1981. Detecting strange attractors in turbulence. Dynam. Systems Turbulence 898, 366–381.

Teager, H.M., Teager, S.M., 1989. Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle, W.J., Marchal (Eds.), Speech Production and Speech Modelling NATO ASI Series D, Vol. 55.

Temam, R., 1993. Infinite-dimensional dynamical systems in mechanics and physics. In: Applied Mathematical Science, Vol. 68. Springer-Verlag, New York.

Thomas, T.J., 1986. A finite element model of fluid flow in the vocal tract. Comput. Speech Language 1, 131–151.

Tokuda, I., Miyano, T., Aihara, K., 2001. Surogate analysis for detecting nonlinear dynamics in normal vowels. J. Acoust. Soc. Am. 110 (6), 3207–3217.

Townshend, B., 1991. Nonlinear prediction of speech signals. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-91, pp. 425–428.

Tritton, D.J., 1988. Physical Fluid Dynamics. Oxford University Press, NY.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2002. The HTK Book, Entropic Ltd.