Filtered Dynamics and Fractal Dimensions for Noisy Speech Recognition

Vassilis Pitsikalis, *Student Member, IEEE*, and Petros Maragos, *Fellow, IEEE*

***Abstract*—We explore methods from fractals and dynamical systems theory for robust processing and recognition of noisy speech. A speech signal is embedded in a multidimensional phase-space and is subsequently filtered exploiting aspects of its unfolded dynamics. Invariant measures (fractal dimensions) of the filtered signal are used as features in automatic speech recognition (ASR). We evaluate the new proposed features as well as the previously proposed multiscale fractal dimension via ASR experiments on the Aurora 2 database. The conducted experiments demonstrate relative improved word accuracy for the fractal features, especially at lower signal-to-noise ratio, when they are combined with the mel-frequency cepstral coefficients.**

***Index Terms*—Automatic speech recognition (ASR), filtered embedding, fractal dimension, phoneme classification.**

I. Introduction

THERE has been strong experimental and theoretical evidence for the existence of important nonlinear aerodynamic phenomena in the vocal tract during speech production [1]. Such phenomena include nonlaminar flow, flow separation, generation and propagation of vortices, and formation of jets. However, the state of the art in acoustic processing for automatic speech recognition (ASR) systems employs features like mel-frequency cepstral coefficients (MFCC) that are based on the linear source-filter model and plane wave propagation in the vocal tract, ignoring nonlinear phenomena of the speech production system. Further, even though several ASR systems have attained satisfactory performance, their efficacy degrades significantly when speech is contaminated with noise [2].

Aerodynamic phenomena observed during speech production indicate the existence of modulations and turbulence that may be generated during formation of various phoneme types [3]. In this letter, we focus on phenomena related to turbulence. It has been conjectured that methods developed in the framework of chaotic dynamical systems and fractal theory might be employed for the analysis of turbulent flow, for example, modeling the multiscale geometrical structures and energy cascades in turbulence by using fractals [4]. Herein, we concentrate on fractal measures as a quantitative characteristic of system complexity. For example, fractal dimension can be interpreted as an approximate quantitative characteristic feature that corresponds to the amount of turbulence that may reside in a speech waveform. Work in this area includes such an application of fractal measures to the analysis of speech signals and recognition [5]. Lately, research in the area has been amplified by ideas concerning state-space reconstruction [6]–[8], which is based on the embedding theorem [9]. In this framework, we consider multidimensional denoising methods inspired by the systems dynamics as alternatives to scalar approaches like filterbank analysis and auditory processing.

In this letter, we address several issues discussed above, and extending our previous work [5], [10], we propose a *combination* of dynamical filtering on embedded noisy speech signals followed by correlation dimension measurements for speech analysis and recognition (see Section III). First, we measure the *correlation dimension* on the reconstructed multidimensional phase-space of speech signals and propose a feature vector containing statistics of the correlation dimension measurements. Further, we conduct broad-type phoneme classification experiments, in order to examine the discriminative ability of the fractal features. Moreover, for the case of noisy speech, we first filter the embedded signal by exploiting its local geometrical structure and subsequently estimate the proposed features. The overall method, namely, filtered dynamics-correlation dimension (FDCD), is evaluated (see Section IV) on Aurora 2 database and attains an average word recognition improvement of 15% over all tests and signal-to-noise ratio (SNR), while for low SNR, the average improvement reaches 37%. Finally, we incorporate the multiscale fractal dimension introduced in [5] in all presented experiments showing positive results.

II. Background

1) Embedding: We assume that the speech production system may be viewed as a nonlinear dynamical system $X(n) \rightarrow X(n+1) = F[X(n)]$. A speech signal $s(n), n = 1, \ldots, N$ is considered a one-dimensional (1-D) projection of a vector function applied to the unknown multidimensional state variables $X(n)$. The embedding vector $Y(n) = [s(n), s(n+T_d), \ldots, s(n+(D_e-1)T_d)]$ formed by samples of the original signal delayed by multiples of T_d defines a motion in a reconstructed D_e-dimensional space. If the unfolding is successful, i.e., the embedding dimension D_e is large enough, according to the embedding theorem [9], the resulting system shall have common invariants, like fractal dimensions, with the original phase-space of $X(n)$.

Low embedding dimensions entail the intersection of distinct system orbits in the reconstructed phase-space, i.e., the system's manifold is folded. In order to determine a sufficient embedding dimension, a true versus false neighbor criterion is formed by comparing the distance between points embedded in successively increasing dimensions; the dimension at which the percentage of false neighbors is minimized is chosen as

Manuscript received August 3, 2005; April 3, 2006. This work was supported in part by the FP6 European research programs HIWIRE and MUSCLE and in part by the "Protagoras" NTUA research program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Frederic Bimbot.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece (e-mail: vpitsik@cs.ntua.gr; maragos@cs.ntua.gr).

Digital Object Identifier 10.1109/LSP.2006.879424

the sufficient embedding dimension. The "optimum" time delay T_d is selected as the location of the first minimum of the average mutual information $I(T) = \sum_{n=1}^{N-T} P(s(n), s(n+T)) \log_2[P(s(n), s(n+T))/(P(s(n))P(s(n+T)))]$, where $P(\cdot)$ is the pdf estimated from the histogram of $s(n)$ [9].

2) Correlation Dimension: The correlation dimension is a type of fractal dimension, corresponds to the number of active degrees of freedom, and indicates the underlying system complexity. It can be practically estimated by employing a method that belongs to the category of average pointwise mass algorithms [11]. The correlation sum C is given for each scale r by the number of points with distance smaller than r, normalized by the number of pairs of points: $C(N,r) = 1/(N(N-1)) \sum_{i=1}^{N} \sum_{j \neq i} U(r - \|X_i - X_j\|)$, where U is the Heavyside unit-step function. Then, the correlation dimension is defined as

$$D_C = \lim_{r \to 0} \lim_{N \to \infty} \log C(N,r)/\log r. \qquad (1)$$

For small enough scales and for large enough N, $C(r) \propto r^{D_C}$.

3) Multiscale Fractal Dimension: The multiscale fractal dimension (MFD) has been proposed for nonlinear speech analysis in [5]. The main concept is based on the morphological covering algorithm that computes the Minkowski–Boulingand dimension D_M of a planar set. This is computed by dilating the graph of the speech signal with disks B of increasing radii ϵ. If $A_B(\epsilon)$ is the area of the dilated graph, D_M equals $2 - \lim_{\epsilon \to \infty} \log[A_B(\epsilon)]/\log(\epsilon)$. This limit can be estimated from the slope of a line fit to the $\log[A_B(\epsilon)]$ versus $\log(\epsilon)$ data using least squares. The successive local estimates of D_M over moving scale windows yield the MFD.

III. Fractal Feature Extraction, Phoneme Classification, and Filtered Dynamics

A. Fractal Features

In the *unfolded* phase-space, we measure C and D_C as in (1) using least-squares local slope estimation of the $\log C(N,r)$ versus $\log r$ data. We form an eight-dimensional correlation dimension feature vector CD by 1) calculating over the whole range of scales r the mean and the variance 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 variances of D_C, in order to include local-scale information.

For the MFD feature set, we estimate D_M on the *scalar* speech signals along the lines of Section II-3 and sample the MFD function at specific scale values ϵ. We have experimentally observed that the variation of the MFD function is better captured by sampling (at six scales) over a logarithmic scale. Both the MFD and the CD features are related to a superset of fractal dimensions [9] and are proposed as distinct tools for complexity-related measurements, for the scalar signals and the multidimensional embedded signals, respectively.

B. Phoneme Classification

In order to explore whether fractal features on their own bear information facilitating the discrimination among broad phoneme classes, we present a set of classification experiments without employing the subsequently presented filtering method. The speech corpus that is utilized is the complete TIMIT database since it is accompanied by hand-labeled phone-level transcriptions. Each signal processed is an *isolated* phoneme. The classification experiments are designed by using the partitioning of phonemes among broad categories. They have been realized using the HTK with one-state HMM for the fractal features and three-state HMM for MFCC using three Gaussian mixtures for each state. MFCC are used as baseline features augmented by derivative and acceleration coefficients.

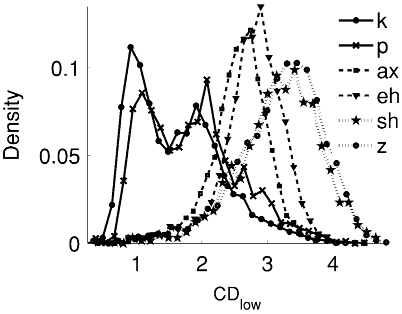

Fig. 1. Histograms of the correlation dimension estimate in the lower scales CD_{low} for selected types of phonemes; number of phonemes: 3394 /k/, 2181 /p/ (stops), 2152 /ax/, 1587 /eh/ (vowels), 5825 /sh/, 3463 /z/ (fricatives).

TABLE I
Classification Scores (%) for Broad Phoneme Classes†
Using the MFCC and the Fractal Features

	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
CD_{low}	65.82	53.60	41.49	43.48
MFD	81.94	71.71	47.50	58.65
	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
CD_{low}	82.35	73.36	77.81	76.42
MFD	88.88	93.40	93.64	88.73

†Classes are vowels (Vo), fricatives (Fr), stops (St), nasals (Na), liquids (Li), voiced (Voi), unvoiced (Un), front (Fro), central (Ce) and back (Ba).

In Fig. 1, we present histograms of a single component of the feature vector CD_{low}, namely, the correlation dimension estimation over the lower scales ($[r_{\min}, \bar{r}]$), for selected phonemes. The specific component stands as a representative that is closest to the correlation dimension estimation and thus of the underlying complexity, since its estimate definition requires that the scale tends to zero. We observe that the estimate of CD_{low} is higher for the fricative sounds and lower for the vowels as intuitively expected, since fricative sounds have more complex dynamics; also, CD_{low} has greater variance for transient sounds like stops than the others. The classification scores for eight experiments shown in Table I indicate the capability of the proposed feature sets to classify phonemes into broad classes; some cases like Voiced versus Unvoiced perform better than Front versus Central versus Back. While the fractal features alone contain six or eight components *per phoneme*, they occasionally yield comparable accuracies to the MFCC feature vector containing 39 coefficients *per frame*.

C. Filtered Dynamics

For noisy speech, we employ a denoising method in the unfolded phase-space. Increased interest has appeared in this field

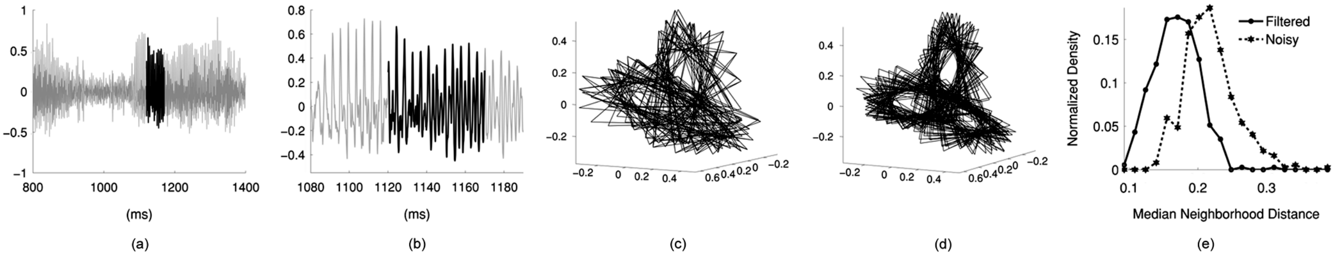

Fig. 2. (a) Noisy speech signal segment (Aurora 2, 0 dB SNR); in dark the processed 50 ms frame, (b) detail. Embedded frame: (c) before and (d) after dynamic filtering; the trajectories in (d) are more compact than in (c). (e) The median distance between a reference point and its 15 nearest neighbors decreases as a result of the filtering procedure: The histogram for all points quantifies the increase in compactness visually observed by comparing (c) and (d).

[12]–[14], while there are a couple of efforts toward the application of such methods to speech signals using limited data sets [7], [15]. In this first *systematic application* on a standard noisy speech database, we adopt features from various methods and evaluate the proposed method on an ASR task. We consider the clean speech signal $s(k)$ contaminated with additive noise $\eta(k)$, which gives the observed signal $s_n(k) = s(k) + \eta(k)$. The embedded signal $Y_n = Y_n(k)$ is corrupted by noise and is thus characterized by increased variance; our objective is to reduce this effect by using a smoothing transformation. The processing is applied separately for each reference point $Y(k)$, using its m (e.g., 30) closest neighbors; these are aggregated in the $m \times D_e$ neighborhood matrix $G(k)$ that shall be henceforth referred to as the neighborhood of $Y(k)$. The geometrical structure of the neighborhood can be represented using a set of eigenvectors estimated by the singular value decomposition of the neighborhood matrix

$$G(k) = U\,S\,V^T. \tag{2}$$

U is a $m \times m$ matrix formed by the eigenvectors of the $G(k)G(k)^T$ matrix, S contains in its diagonal the singular values in decreasing order, and V is a $D_e \times D_e$ matrix formed by the eigenvectors of the covariance matrix $G(k)^T G(k)$.

If the noise variance is lower than that of the clean signal, the larger eigenvalues will correspond to the system dynamics, whereas the smaller ones will correspond to noise. To suppress the latter, we project the data on the subspace spanned by the p largest principal components that account for a fixed factor λ (e.g., 0.7) of the total variance of the set $G(k)$, i.e., $\sum_{i=1}^{p} \sigma_i^2 = \lambda \sum_{i=1}^{N} \sigma_i^2$. We apply this projection step only to the central reference point, so as not to distort points that are on the neighborhood boundaries. The proposed change in the location of each point is contracted by a factor, which is linearly increased along with the iterations, in the range [0.1, 1], in order to move the points gradually toward the positions suggested by the geometry of their neighborhood. This projection step is repeated for the whole set since the neighborhoods of each point may be altered in each iteration, and only a part of the whole correction suggested is applied. Instant application of the whole correction could lead to instabilities. This process is applied until convergence or for a fixed number of iterations (e.g., 8–12). Further robustness is acquired by applying a lowpass filter as in [12] to the scalar noisy signal $s_n(k)$ before the embedding. However, the corrections that are computed via the local projections are applied to the original noisy signal. Fig. 2 shows the effect of the projection-based cleaning procedure on the embedded frame. Visually, the trajectories in various regions are more compact, and the points are more condensed according to the geometry of the dynamics when compared to the noisy one. Quantitatively, we observe [see Fig. 2(e)] the reduction of the nearest neighborhood distances, by measuring the median distance between each reference point and a fixed number of its nearest neighbors.

IV. Application to Speech Recognition

The overall method proposed above for noisy speech signals, namely, *filtered dynamics-correlation dimension* (FDCD), is evaluated via ASR experiments in Aurora 2 database [2], which contains additive noise in various conditions and SNR. A drawback of the FDCD method is its high computational complexity, due to the intensive embedding and filtered embedding procedures (computation time is up to two or three orders of magnitude higher than the one for the MFCC, respectively). The experiments are realized by use of the HTK system (context-independent, 18-state, left–right word HMM with three Gaussian mixtures). The HMM are trained on the clean data set (8440 utterances) and tested on 60 noisy sets (6 SNR × 10 noise-types × 1001 utterances) referred to as *clean training* scenario (we focus on this scenario that examines the mismatch among clean and noisy data on various SNR and noise types). The above nonlinear features are not self-standing but contribute as second-order information to the first-order linear speech structure expressed by MFCC. The average word recognition accuracy of the fractal features *alone* ranges between 61% and 26% for the clean and 10 dB SNR cases, respectively. The input vectors are split in two data streams that are assumed independent with stream weights set equal to 1 and 0.2 for MFCC and fractal features, respectively. The augmented features include 13 elements for the MFCC plus normalized energy augmented by six or eight elements for the MFD or FDCD features, respectively. All feature vectors are extended by their time derivatives $(\Delta, \Delta\Delta)$, and cepstral mean subtraction is applied to increase their robustness. The frame length for the MFCC and MFD streams is set equal to 30 ms. For the FDCD stream, additional information surrounding each frame is considered using 50 ms frames; since dynamical processing prompts for longer time series, selecting shorter frame length deteriorates slightly the results (e.g., by 0.5% for 40 ms). Synchronization between the different streams is achieved by keeping the same 10-ms frame period.

Table II shows the average results per SNR (sets A, B contain four noise types each, and set C contains two). The combination of MFCC with the MFD features results in a slight improvement for higher SNR that increases for the middle SNR. Although the dynamical filtering is proposed prior to the CD measurements,

TABLE II
AVERAGE (AVG.) WORD ACCURACY (%) IN ALL TESTS AND AVERAGE RELATIVE IMPROVEMENT (%) (IMPROV.) OF THE MFCC AND THE AUGMENTED FRACTAL FEATURES ON AURORA 2 (CLEAN TRAINING)[‡]

	SNR	clean	20 dB	15 dB	10 dB	5 dB	0 dB
Set	Feature						
(A)	MFCC	98.71	95.67	88.73	68.91	38.44	15.67
	+MFD	98.74	96.42	91.66	77.75	49.38	18.68
Avg.	Improv.	0.03	0.79	3.31	12.82	28.47	19.21
	+FDCD	98.60	96.46	92.71	82.93	58.66	19.55
	Improv.	-0.11	0.83	4.48	20.35	52.60	24.77
(B)	MFCC	98.71	96.24	88.79	69.84	42.03	17.69
	+MFD	98.74	96.98	91.99	78.80	51.99	20.92
Avg.	Improv.	0.03	0.77	3.61	12.83	23.69	18.24
	+FDCD	98.62	96.82	93.20	83.26	59.48	23.18
	Improv.	-0.10	0.61	4.97	19.21	41.52	31.01
(C)	MFCC	98.58	95.19	89.64	75.51	49.87	24.68
	+MFD	98.72	95.70	91.28	80.80	57.02	25.49
Avg.	Improv.	0.15	0.54	1.82	7.01	14.33	3.32
	+FDCD	98.51	95.74	92.26	82.56	58.93	24.14
	Improv.	-0.07	0.57	2.92	9.34	18.16	-2.17
Global Avg.	MFCC	98.66	95.70	89.05	71.42	43.45	16.68
	+MFD	98.73	96.37	91.64	79.11	52.80	21.70
	Improv.	0.07	0.70	2.91	10.88	22.16	13.59
	+FDCD	98.58	96.34	92.72	82.92	59.02	22.29
	Improv.	-0.09	0.67	4.12	16.30	37.43	17.87

[‡] Compared to the 61.1% performance of [2] the average improvement over all tests (0 - 20 dB SNR) is : +FDCD: 15.7%, +MFD: 11.9%.

TABLE III
AVERAGE WORD ACCURACY (%) OVER ALL TESTS OF PLAIN CD AND FDCD ON AURORA 2 (CLEAN TRAINING)

SNR	20 dB	15 dB	10 dB	5 dB	0 dB
Feature					
MFCC+plain CD	96.58	92.77	79.49	46.98	14.97
MFCC+FDCD	96.34	92.72	82.92	59.02	22.29

by projecting back to the 1-D signal, one may employ it for enhancement and further estimate the MFD feature. This experiment has similar performance in most SNR, with a light improvement relative to the MFD case without filtering (up to 2.4% at 5 dB SNR); anyhow, the combination of filtering and MFD has inferior performance compared to FDCD. Higher improvements are shown for the FDCD features (overall improvement MFD: 10.1%, FDCD: 15.3%; 5 dB SNR improvement MFD: 22.2%, FDCD: 37.4%), except for the clean case, where the MFD features perform slightly better (the utilized baseline typically performs better than the one reported in [2]). Compared to these results, the average performance of the recently proposed modulation features [16] is similar or slightly better on average on Aurora 2 (average overall improvement 21%; at 5 dB SNR: average 33%, maximum 46%). However, the modulation features extract different types of nonlinear information than the fractal features. Finally, Table III indicates that the improved ASR performance attained by the CD features is in part due to the filtering process: when using the unfiltered data points, the average overall improvement decreases. This demonstrates that the filtered dynamics step is essential for the feature extraction of noise-robust features that further incorporate information related to the system's invariants.

V. CONCLUSION

Based on a dynamical systems perspective, we employ fractal dimension measurements on a speech signal's multidimensional phase-space. Our goal is to quantify turbulence-related phenomena and to cope with additive noise. The fractal features (CD and the previously proposed MFD) are shown to possess discriminative ability for the classification among broad classes like fricatives, vowels, voiced, and unvoiced phonemes from the TIMIT database. Moreover, the combination of filtering on the embedded space followed by the correlation dimension feature extraction (FDCD) results in an average relative improvement of 37% for recognition at lower SNRs on the Aurora 2 database, when the extracted features are combined with the MFCC. This indicates the potential of the dynamical systems approach and of fractal dimension measurements to robustly extract nonlinear information useful for ASR.

REFERENCES

[1] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech Prod. Speech Model., NATO ASI Ser. D*, vol. 55, 1989.

[2] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, Paris, France, Sep. 2000.

[3] P. Maragos, A. G. Dimakis, and I. Kokkinos, "Some advances in nonlinear speech modeling using modulations, fractals and chaos," in *Proc. DSP*, Santorini, Greece, Jul. 2002.

[4] B. Mandelbrot, *The Fractal Geometry of Nature*. San Francisco, CA: Freeman, 1982.

[5] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1925–1932, 1999.

[6] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-domain isolated phoneme classification using reconstructed phase spaces," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 458–466, Jul. 2005.

[7] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 1–17, Jan. 1999.

[8] S. Narayanan and A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants," *J. Acoust. Soc. Amer.*, vol. 97, no. 4, pp. 2511–2524, 1995.

[9] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*. New York: Springer-Verlag, 1996.

[10] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," in *Proc. ICASSP*, Orlando, FL, 2002.

[11] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D*, vol. 9, pp. 189–208, 1983.

[12] T. Sauer, "A noise reduction method for signals from nonlinear systems," *Physica D*, vol. 58, pp. 193–201, 1992.

[13] R. Cawley and G. H. Hsu, "Local-geometric projection method for noise reduction in chaotic maps and flows," *Phys. Rev. A*, vol. 46, pp. 3057–3082, 1992.

[14] A. G. Darbyshire and D. S. Broomhead, "Robust estimation of tangent maps and Lyapunov spectra," *Physica D*, vol. 89, pp. 287–305, 1996.

[15] R. Hegger, H. Kantz, and L. Matassini, "Denoising human speech signals using chaoslike features," *Phys. Rev. Lett.*, vol. 84, no. 14, pp. 3197–3200, 2000.

[16] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621–624, Sep. 2005.