MODULATION AND CHAOTIC ACOUSTIC FEATURES FOR SPEECH RECOGNITION *

Dimitrios Dimitriadis†, Petros Maragos†, Vasilis Pitsikalis† and Alexandros Potamianos‡

† Dept. ECE, National Technical University of Athens, Zografou, 15773 Athens, Greece
‡ Bell Laboratories, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.

October 3, 2001

Abstract

Automatic speech recognition systems can benefit from including into their acoustic processing part new features that account for various nonlinear and time-varying phenomena during speech production. In this paper, we develop robust methods for extracting novel acoustic features from speech signals based on nonlinear and time-varying models of speech. These new modulation- and chaotic-type features are integrated with the standard linear ones (mel-frequency cesptrum) to develop a generalized hybrid set of acoustic features. The efficacy by showing significant improvements in HMM-based phoneme recognition over the TIMIT database.

Key Words: non-linear, modulation, chaotic, recognition, ASR.

Accepted in *Journal of Control and Intelligent Systems*
Special Issue on Nonlinear Speech Processing

*This research work was supported by the Greek Secretariat for Research and Technology and by the European Union under the program ΕΠΕΤ-98 with Grant # 98ΓΤ26. It was also partially supported by the basic research program ARCHIMEDES of the NTUA Institute of Communication and Computer Systems. D. Dimitriadis, P. Maragos and V. Pitsikalis are with the National Technical University of Athens, Dept. of Electrical and Computer Engineering, Zografou, Athens 15773, Greece. E-mail: [ddim,maragos,vpitsik]@cs.ntua.gr. A. Potamianos is with Bell Laboratories, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A. E-mail: potam@research.bell-labs.com

1 Introduction

Despite many decades of research, the current automatic speech recognition (ASR) systems are still inferior to the corresponding human cognitive abilities because of the many limitations of their acoustic processing, pattern recognition and linguistic subsystems. Thus, both from a scientific and a technology viewpoint, there is a significant interest in improving ASR systems. For several decades the traditional approach to speech modelling has been the linear (source-filter) model where the true nonlinear physics of speech production is approximated via the standard assumptions of linear acoustics and 1D plane wave propagation of the sound in the vocal tract. The linear model has been applied to speech coding, synthesis and recognition with limited success [1, 2]; to built successful applications, deviations from the linear model are often modeled as second-order effects or error terms. There is indeed strong theoretical and experimental evidence [3, 4, 5, 6] for the existence of important nonlinear aerodynamic phenomena during the speech production that cannot be accounted for by the linear model. The investigation of speech nonlinearities can proceed in at least two directions: (i) numerical simulations of the nonlinear differential (Navier-Stokes) equations governing the 3-D dynamics of the speech airflow in the vocal tract, and (ii) development of nonlinear signal processing systems suitable to detect such phenomena and extract related information. In our research we focus on the second approach, which is computationally much simpler, i.e., to develop models and extract related acoustic signal features describing two types of nonlinear phenomena in speech, *modulations* and *turbulence*. These novel features are then applied to speech recognition.

The traditionally applied "standard" speech features used in ASR are based on short-time smoothed cepstra stemming from the linear model. This representation ignores the nonlinear aspects of speech and is sensitive to small signal durations. Adding new robust nonlinear information is however quite promising to lead to improved performances and robustness. In this paper, we focus on improving the acoustic processing part of ASR systems by developing robust nonlinear and instantaneous features based on modulation and chaotic models for speech production and by using these features to increase the recognition performance of ASR systems whose pattern classification part is based on Hidden Markov Models (HMM). Our motivation for this research work includes the following: (1) In prior work [7, 8, 9] some of the authors have shown that the AM-FM modulation model and instantaneous demodulation algorithms for speech resonances can track nonstationarity in speech and lead to better performance in several speech applications. Some preliminary work on using Teager energy features (that indirectly contain pre-modulation information) in speaker and speech recognition include [10, 11, 12, 13]. (2) By using concepts from fractals [14] to quantify the geometrical roughness of speech waveforms, some of the authors were able to extract fractal features from speech signals and use them to improve phonemic recognition [15]. (3) Fractals can quantify the geometry of speech turbulence. A fuller account of the nonlinear dynamics can be obtained by using chaotic models for general time-series as in [16].

Section 2 of this paper reviews the use of modulation models for speech resonances and describes robust demodulation algorithms for extracting the parameters of such models. Section 3 summarizes the basic concepts and algorithms for analyzing speech signals with chaotic models. In Section 4 we describe how to extract novel short-time feature vectors from speech signals that contain modulation and/or chaotic dynamics information, integrate these nonlinear speech features with the standard linear ones (cepstrum), and develop a generalized set of acoustic features for improving HMM-based phonemic recognition.

2 Speech Modulation Model and Demodulation Algorithms

By 'speech resonances' we shall loosely refer to the oscillator systems formed by local vocal tract cavities emphasizing certain frequencies and de-emphasizing others. Although the linear model assumes that each speech resonance signal is a damped cosine with constant frequency within 10-30 ms and exponentially decaying amplitude, there is much experimental and theoretical evidence for the existence of *amplitude modulation (AM) and frequency modulation (FM)* in speech resonance

signals, which make the amplitude and frequency of the resonance vary instantaneously within a pitch period. Motivated by this evidence, Maragos, Quatieri and Kaiser [17, 7] proposed to *model each speech resonance with an AM-FM signal*

$$x(t) = a(t)\cos[2\pi \int_0^t f(\tau)d\tau] \tag{1}$$

and the total speech signal as a superposition of such AM-FM signals, one for each formant. Here $a(t)$ is the instantaneous amplitude signal and $f(t)$ is the instantaneous frequency representing the time-varying formant signal. The short-time formant frequency average $f_c = (1/T)\int_0^T f(t)dt$, where T is in the order of a pitch period, is viewed as the carrier frequency of the AM-FM signal. The classical linear model of speech views a formant frequency as constant, i.e., equal to f_c, over a short time (10-30 ms) frame. However, the AM-FM model can both yield the average f_c and provide additional information about the formant's instantaneous frequency deviation $f(t) - f_c$ and its amplitude intensity $|a(t)|$. To isolate a single resonance from the original speech signal, bandpass filtering is first applied around estimates of formant center frequencies. Then for demodulating a resonance signal, Maragos et al. [7] used the nonlinear Teager-Kaiser energy-tracking operator

$$\Psi[x(t)] \triangleq \left[\frac{dx(t)}{dt}\right]^2 - x(t)\frac{d^2x(t)}{dt^2} \tag{2}$$

to develop the following nonlinear algorithm

$$\frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx f(t) \quad , \quad \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)| \tag{3}$$

This is the *energy separation algorithm (ESA)* and provides AM-FM demodulation by tracking the physical energy implicit in the source producing the observed acoustic resonance signal and separating it into its amplitude and frequency components. It yields very good estimates of the instantaneous frequency signal $f(t) \geq 0$ and of the amplitude envelope $|a(t)|$ of an AM-FM signal, assuming that $a(t), f(t)$ do not vary too fast (small bandwidths) or too greatly compared with the carrier frequency f_c. There is also a very efficient and computationally simple *discrete* version of the ESA, called *DESA* [7], which is obtained by using a discrete energy operator on discrete-time nonstationary sinusoids. The DESA is a novel and very promising approach to demodulating speech resonances for many reasons: (i) It yields very *small errors* for AM-FM demodulation. (ii) It has an extremely *low computational complexity*. (iii) It has an excellent time resolution, almost *instantaneous*; i.e., operates on a 5-sample moving window and can track instantaneous changes of speech modulations.

Extensive experiments on speech demodulation using the DESA in [7, 8, 9] indicate that these amplitude/frequency modulations *exist* in real speech resonances and are necessary for its *naturalness*.

The main disadvantage of the DESA is a moderate sensitivity to noise. Thus, we describe next an alternative approach [18] where we first interpolate the discrete-time signal using smoothing splines [19], and then apply the continuous-time ESA (3). *Splines* are piecewise polynomial functions constructed as a linear combination of B-Splines. A spline function of order ν has continuous derivatives up to order $\nu - 1$, which is important when using the energy operator Ψ. At first we used exact splines to improve the performance of the ESA, tested on noisy AM-FM signals with different levels of SNR. The results were disappointing as the exact fitting of the curve, due to the presence of noise, was creating large estimation errors. The problem of noise led us to optimally interpolate signal samples with *smoothing splines*, whose main advantage is that the interpolating polynomial does not pass through the signal samples but close enough. The smooth spline interpolating function is defined as the function s_ν that minimizes the mean square error criterion

$$E = \underbrace{\sum_{n=-\infty}^{+\infty} (x[n] - s_\nu(n))^2}_{E_d} + \lambda \underbrace{\int_{-\infty}^{+\infty} \left(\frac{\partial^r s_\nu(t)}{\partial t^r}\right)^2 dt}_{E_s}$$

where E_d is the data fitting error and E_s quantifies the non-smoothness ("roughness") of the interpolant by the mean square value of its derivative. The positive design parameter λ controls the trade-off between how smooth the interpolating curve will be and how close to the data points the interpolant will pass. (For $\lambda = 0$ we obtain exact splines with no data smoothing.) Given the initial signal samples $x[n],\ n = 1, \ldots, N$, the interpolating spline function of order $\nu = 2r - 1$ is given by [19]

$$s_\nu(t) = \sum_{n=-\infty}^{+\infty} c[n]\beta_\nu(t-n) \tag{4}$$

where $\beta_\nu(t)$ is the B-spline of order ν, and the coefficients $c[n]$ depend only on the data $x[n]$, the parameter λ and the analytic expression of the B-spline. The coefficient sequence $c[n]$ can be determined recursively by using the sequence $x[n]$ as input to excite an IIR filter with transfer function $H_\nu^\lambda(z) = 1/[B_\nu(z) + \lambda(-z + 2 - z^{-1})^{\frac{\nu+1}{2}}]$, where $B_\nu(z)$ is the Z-transform of the discrete spline $b_\nu[n] = \beta_\nu(n)$. The IIR filter H_ν^λ has a symmetric impulse response, and all its poles are inside the unit circle. Thus, the spline coefficients $c[n]$ can be determined stably via a few recursive equations [18, 19].

The above spline interpolation leads us to a new approach for ESA-based demodulation whose basic steps are the following. (i) By using smoothing splines, the original discrete-time signal $x[n]$ is interpolated to create a continuous-time expansion $s_\nu(t)$. For fixed ν, λ, the interpolation procedure is complete after the computation of the coefficient sequence $c[n]$. (ii) The continuous-time energy operator Ψ and the continuous ESA are applied to the continuous-time signal $s_\nu(t)$. This requires computing $\Psi[s_\nu(t)]$ and $\Psi[\partial s_\nu(t)/\partial t]$, which in turn require the derivatives $\partial^r s_\nu(t)/\partial t^r$ for $r = 1, 2, 3$. We can derive closed-form expressions for these derivatives that involve only the coefficients $c[n]$ and the B-spline functions [18]. For example,

$$\frac{\partial s_\nu(t)}{\partial t} = \sum_n \left(c[n] - c[n-1]\right) \beta_{\nu-1}(t - n + 1/2) \tag{5}$$

The continuous ESA (3) can estimate the instantaneous amplitude $a(t)$ and frequency $f(t)$ of the continuous signal $s_\nu(t)$. (iii) The information-bearing signals $a(t), f(t)$ are sampled to obtain estimates of the instantaneous amplitude $A[n] = a(nT)$ and frequency $F[n] = Tf(nT)$ of the original discrete signal $x[n]$. This whole approach above is called the **Spline-ESA**.

By setting $\nu = 5$, the time-window (i.e., the number of input samples required to produce one output sample) of Spline-ESA becomes the same with that of the DESA. Extensive comparisons [18] between the Spline-ESA (with $\nu = 5$ and λ fixed to a constant value in the order of 0.25) versus the DESA have demonstrated that, while both algorithms perform well in signal-plus-noise environments with high SNRs, the Spline-ESA outperforms the DESA in low SNRs. This robustness in the presence of noise is the main advantage of the Spline-ESA.

The ESAs are efficient demodulation algorithms only when they are used on narrowband AM-FM signals [20]. This constraint makes the use of *filterbanks* (i.e., parallel arrays of bandpass filters) inevitable for wideband signals like speech. Thus, each short-time segment (analysis frame) of a speech signal is simultaneously filtered by all the bandpass filters of the filterbank, and then each filter output is demodulated using the ESA. In our on-going research on speech analysis and recognition [8, 13] we have been using filterbanks with Gabor bandpass filters whose center frequencies are spaced either linearly or on a mel-frequency scale. Figure 1 shows an example of demodulating three bands of a speech phoneme into their instantaneous amplitude and frequency signals.

3 Speech Analysis using Chaotic Models

Many speech sounds, especially fricatives and plosives, contain various amounts of turbulence. In the linear speech modelling this has been dealt with by having a white noise source exciting the vocal tract filter. It has been conjectured that geometrical structures in turbulence can be modeled

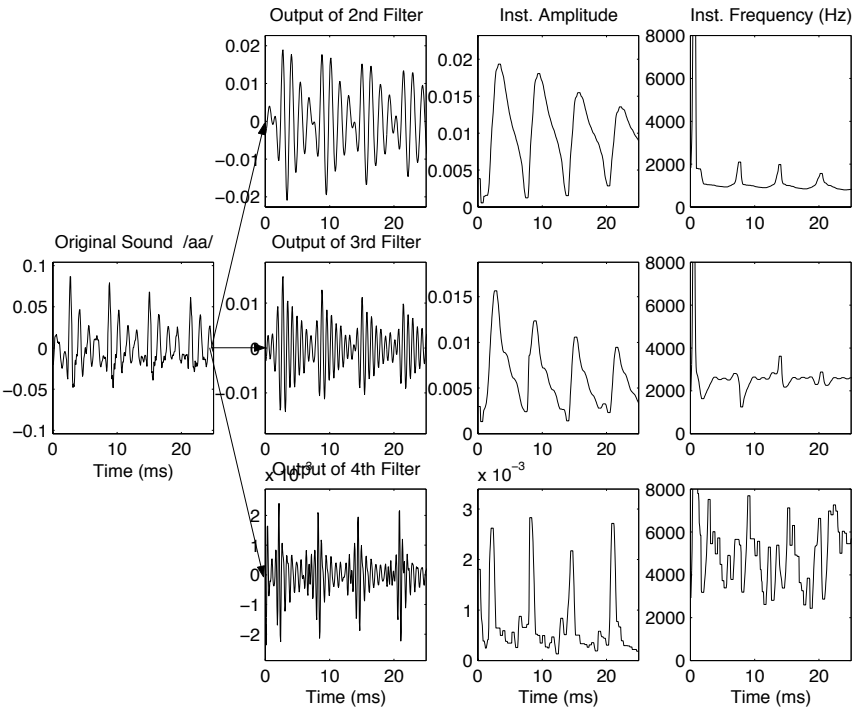

Figure 1: Demodulating a speech phoneme using a Gabor filterbank and the Spline-ESA.

using fractals [14], while its dynamics can be modeled using the theory of chaos. In previous work [15], some of the authors measured the *short-time fractal dimension* of speech sounds as a feature to approximately quantify the degree of turbulence in them and used it to improve phoneme recognition. In this paper, we shall use concepts from chaos theory [16] to model the nonlinear dynamics in speech of the chaotic type.

We assume that (in discrete time n) the speech production system (whose aerodynamics are governed by the 3D Navier-Stokes equations) can be viewed as a nonlinear (but finite dimensional [21] due to dissipativity) dynamical system $X(n) \to F[X(n)] = X(n+1)$ where the phase space of $X(n)$ is multidimensional. A speech signal segment $s(n)$, $n = 1, ..., N$, can be considered as a 1D projection of a vector function applied to the unknown dynamic variables $X(n)$. It is possible that the complexity or randomness observed in the scalar signal could be due to loss of information during the projection. According to the *embedding* theorem [16], the vector

$$Y(n) = [s(n), s(n+T_D), s(n+2T_D), \ldots, s(n+(D_E-1)T_D] \tag{6}$$

formed by samples of the original signal delayed by multiples of a constant time delay T_D defines a motion in a reconstructed D_E-dimensional space that has many common aspects with the original phase space of $X(n)$. Specifically, many quantities of the original dynamical system (e.g. generalized fractal dimensions and Lyapunov exponents) in the original phase-space $X(n)$ are conserved in the reconstructed space traced by $Y(n)$. Thus, by studying the constructible dynamical system $Y(n) \to Y(n+1)$ we can uncover useful information about the original unknown dynamical system $X(n) \to X(n+1)$ provided that the unfolding of the dynamics is successful, e.g. the embedding dimension D_E is large enough. However, the embedding theorem does not specify a method to determine the required parameters (T_D, D_E) but only sets constraints on their values. For example, D_E must be greater than the box-counting dimension of the attractor set. And T_D can have any value except from $p\Delta t$, where $p = 1, 2$ and Δt corresponds to the period of periodic orbits of the system. Hence, procedures to estimate the values of these parameters are essential. The time delay corresponds to the constant time difference between the neighboring elements of each reconstructed vector. The smaller T_D gets, the more will the successive elements be correlated, as not enough time will have elapsed for the system to generate sufficient amounts of information and all connected variables affect the observed one. On the contrary, the greater T_D gets, the more random will the

successive elements be. Thus it is necessary to compromise between these two conflicting arguments. To achieve this, the following measure of nonlinear correlation is used for dealing with chaotic data $s(n)$ [16]:

$$I(T) = \sum_{n=1}^{N-T} P(s(n), s(n+T)) \cdot \log_2 \left[\frac{P(s(n), s(n+T))}{P(s(n))P(s(n+T))} \right] \tag{7}$$

where $P(\cdot)$ denotes probability. Each log term in the above sum is the mutual information for a pair of observed values $s(n), s(n+T)$ which are apart from each other by a delay T. If these values are independent, their mutual information is zero. Thus, $I(T)$ is the *average mutual information* between pairs of samples of the signal segment that are T positions apart. Then, the 'optimum' time delay T_D is selected as the smallest T at which the average mutual information assumes a minimum value:

$$T_D = \min\{\arg\min_{T \geq 0} \ I(T)\} \tag{8}$$

After setting T_D, the next step is to select the dimension D_E of the reconstructed vectors. As a consequence of the projection, points of the 1D signal are not necessarily in their relative positions because of the true dynamics of the multidimensional system (true neighbors). A true vs. false neighbor criterion is formed by comparing the distance between two points S_n, S_j embedded in successive increasing dimensions. If their distance $d_D(S_n, S_j)$ along dimension D is significantly different that their distance $d_{D+1}(S_n, S_j)$ along dimension $D+1$, then they are considered to be a pair of *false neighbors*. Equivalently, if $\frac{d_{D+1}(S_n,S_j) - d_D(S_n,S_j)}{d_D(S_n,S_j)}$ exceeds a threshold (usually in the range of $[10, 15]$), then the two points are false neighbors. The dimension D along which the percentage of false neighbors goes to zero (or minimized in the existence of noise) is chosen as the embedding dimension D_E.

In the unfolded state-space one can measure invariant quantities of the attractor, which if chaotic would be characterized by sensitive dependence on initial conditions, dense periodic points and mixing [22], such as fractal dimensions of geometrical (e.g. box-counting dimension) and/or probabilistic (e.g. information dimension) character. The dimension of the attractor except from being a measure of complexity, corresponds to the number of active degrees of freedom of the system. The *correlation dimension* [23, 22] (belonging to a greater set of generalized dimensions of probabilistic type) is defined as

$$D_C = \lim_{r \to 0} \lim_{N \to \infty} \frac{\log C(N, r)}{\log r}, \tag{9}$$

where C is the correlation sum i.e. for each scale r the number of points with distances less than r normalized to the number of pairs of points:

$$C(N, r) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \theta(r - \|X_i - X_j\|) \tag{10}$$

where θ is the Heavyside unit-step function. Figure 2 shows the waveforms of two speech phonemes, their attractors and correlation dimension measurements. The shape[1] differences in the two attractors are consistent to the corresponding physics for each phoneme.

4 Nonlinear Feature Extraction and Phoneme Recognition

The feature vectors used in speech recognition are typically computed over a 20-30 ms window and are updated every 5-10 ms. The 'standard' feature set consists of the mean square amplitude

[1]The visualization of the multidimensional attractors has been done by showing the first three elements of each vector in 3D space and the last three as RGB color components.

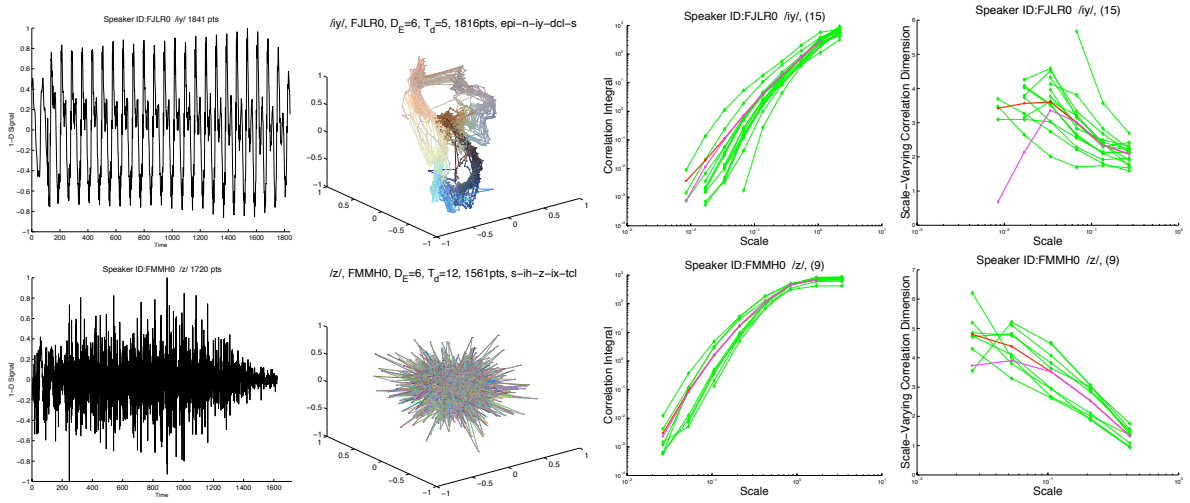

Figure 2: (a) Speech Waveforms, (b) Attractors of Embedded Signals, (c) Correlation Sums, (d) Scale-Varying Correlation Dimensions. Top row: vowel /iy/, bottom row: fricative/z/. (In (c) and (d) thick lines show average curves.)

(usually called 'energy'[2]), the first twelve *mel-frequency cepstrum coefficients (MFCC)* and their first and second time derivatives.

We shall augment the 'standard' feature vector and thus create a *hybrid feature vector* by incorporating information from the non-linear structure of speech of the modulation and chaotic type as additional features. Thus, as short-time acoustic representations of speech we use feature vectors that contain information both from the smoothed cepstrum of the linear model, which represents a first-order approximation to the true speech acoustics, as well as from the speech modulations and the chaotic dynamics, which contain information from the second-order non-linear speech acoustics.

We have used the hybrid feature vector as input to a hidden Markov model (HMM)–based speech recognizer. The HMM recognizer is the HTK system [24]. In the experiments presented below, context-independent 5-state left-right phone HMMs were used. The input vectors are split into different data streams, one for the standard features (MFCC) and the others for the non-linear features. The non-linear features are assumed to be independent of the linear features and to belong to separate probability 'streams'. Each one of these streams has an independent probability distribution. These distributions are modelled by a certain number of Gaussian mixture probability densities, called mixture components. For more details see [24].

We have experimented with a broad range for the number of Gaussian mixture densities, but here we are presenting the recognition results only for the cases of 8 and 16 mixtures, since these values are the most representative. Stream-weights affect directly the recognition process. Our experiments have shown that the best recognition results are obtained when the data-stream weights are equal and sum up to one. So, when using two data-streams (i.e. linear and modulation features or linear and chaotic features) we set each of the weights equal to 0.5, whereas in the case of three data-streams (i.e. linear, modulation and chaotic features) we set the weights equal to 0.35.

The experiments were made over the TIMIT database. The TIMIT database consists of 6300 sentences, i.e. 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the US. All of the speech signals in TIMIT are sampled at 16 kHz. The training set consists of 3696 sentences and the test set of 1344 sentences. Each one of these sentences was segmented into 25-ms speech frames, whose update period was 10 ms. The (linear and nonlinear) feature sets were extracted from each such frame.

[2]We prefer the term 'mean square amplitude' over the term 'energy' because the energy in an oscillatory signal is more appropriate to be related to the physical energy of the source producing this signal. Such an energy is proportional both to the amplitude squared and the frequency squared.

4.1 Modulation Features for Speech Recognition

We have automated the extraction of modulation features from speech signals in the following way: First, we use a parallel filterbank of overlapping Gabor bandpass filters whose center frequencies are spaced on a mel-frequency scale. Second, the output signals from each Gabor bandpass filter are demodulated via the Spline-ESA into its instantaneous amplitude $a(t)$ and frequency $f(t)$ component signals. These lowpass information signals are segmented into 25-ms frames, updated every 10 ms. For each such short-time analysis frame and for each band, the weighted mean F_w and standard deviation B_w of the instantaneous frequency signal are estimated as in [8]:

$$F_w \triangleq \frac{\int_{t_0}^{t_0+T} f(t)a^2(t)dt}{\int_{t_0}^{t_0+T} a^2(t)dt}, \quad B_w \triangleq \frac{\int_{t_0}^{t_0+T} [(\dot{a}(t)/2\pi)^2 + (f(t)-F_w)^2 a^2(t)]dt}{\int_{t_0}^{t_0+T} a^2(t)dt} \tag{11}$$

where t_0 and T are the start and duration of the analysis frame, respectively. Next, we compute the *frequency modulation (FM) percentage* in each band as the ratio $K = B_w/F_w$. For each analysis frame, the FM percentages K_i, $i = 1, ..., L$, are computed, one for each narrowband speech component, where L is the number of filters in the filterbank. The modulation feature set consists of the sequence of the FM percentages K_i and their first and second time derivatives. This is a total of $3L$ numbers per frame. We have experimented with mel-spaced filterbanks consisting of $L = 12$ and $L = 6$ Gabor filters spanning the whole frequency range and overlapping by 50%.

We have used these modulation feature vectors to augment the standard feature vectors employed in speech recognition tasks. The HTK system [24] was used both as the HMM recognizer and for the extraction of the standard feature set which consists of the first 12 mel-scale cepstrum coefficients, the signal's mean-square amplitude and their first and second time-derivatives. So, the standard feature vector's size is 39. The augmented hybrid feature set consists of the standard and the modulation feature set. The two different feature subsets are treated as separate streams (with weights 0.5 each) by the HTK system and their probability distributions are assumed independent.

Table 1: Recognition Results

Phoneme Percent Correct [3]				
# Gaussian Mixtures	MFCC	MFCC+FM	MFCC+Chaotic	MFCC+FM+Chaotic
8	73.95	84.31	78.61	84.75
16	78.76	86.83	85.01	87.69

Table 1 reports the phoneme recognition results over the TIMIT database using either only the standard features (column MFCC) or the augmented standard-plus-modulation features (column MFCC+FM). Clearly, our experiments on phoneme-recognition by augmenting the standard feature set with modulation information, show a significant improvement over using only the standard features. This relative error rate reduction approaches 40% when using 8 Gaussian mixtures. In general, the absolute recognition scores improve with the number of Gaussian mixtures used. Thus, the FM modulation percentage features provide an improvement to the recognition performance with a moderate increase in the size of the feature vector.

The results in table 1 refer to the case of a 6-channel filterbank (i.e. 18 modulation features); hence, the augmented feature set has a size of 57. We have experimentally found that measuring the modulations in the outputs of only 6 Gabor filters yields better recognition results than using 12 filters. For example, the correct phoneme recognition for the 12-channel filterbank was 80.96% (using 8 Gaussian mixtures) compared to 84.3% for 6 channels. Note that the 12-channel case employs a larger feature vector of size 75 despite its inferior recognition performance. This difference in the recognition rates can be explained based on the modulation model for speech resonances. In the 12-channel case the large number of filters causes each bandpass filter to have a narrower

[3] The percentage number of phonemes correctly recognized is given by the ratio of the number of correct labels to the total number of phonemes in the defining transcription files.

bandwidth and hence pass a smaller part of the AM-FM modulation structure of the neighbor speech resonances. In contrast, the filters in the 6-channel filterbank have a wider bandwidth and hence they keep a richer part of the modulation information.

4.2 Chaotic Features for Speech Recognition

As explained in Section 3, through an automated procedure each speech analysis frame has been embedded in a multidimensional state-space using the appropriate time delay T_D and embedding dimension D_E. The physical justification of embedding only a frame instead of a whole phoneme is that the reconstructed space in this occasion belongs to the state-space of the dynamic system during the time period it produced the current frame. After the embedding, we computed a feature vector that was related to the correlation sum and the scale-varying correlation dimension and hence carried information about the chaotic dynamics of each frame. Specifically, we selected a set of four chaotic features: (1) the mean of the correlation sum C, (2) the standard deviation of C, (3) the mean of the scale-varying correlation dimension D_C, and (4) the standard deviation of D_C. This feature set also included the first and second time derivatives of these four features.

We have used the above chaotic feature set to augment the standard feature set (MFCC) and test it on HMM-based recognition over the TIMIT database, experimenting with a wide range of stream weights and number of Gaussian mixtures $(1 - 16)$. The recognition results of the hybrid feature set (MFCC + Chaotic) were quite promising, even though our preliminary first application of chaotic features in an ASR system used the fewest and simplest possible such features. One of the best recognition percentages (see table 1) resulted with weights $(0.5, 0.5)$ for the corresponding streams (standard, chaotic) and for a 16-mixture model. The relative phone error rate reduction of 29% (over using only the standard features) is significant and is possibly due to the detection of nonlinear phenomena which remain "hidden" in the 1D dynamics. Unfolding the signal to its original state-space enables the observation of the true dynamics of the system; furthermore a broad variety of new measurements can be performed on the unfolded attractor that can yield fractal and/or chaotic features.

Of special interest is the experiment reported in the rightmost column of table 1 in which both the chaotic features and the modulation features were used to augment the standard feature set. This produced a hybrid feature vector of dimension 69. Using equal weights of 0.35 for the three data streams (standard, modulation, chaotic) outperformed all other experiments, achieving relative error rate reduction by 42% for both 8 mixtures and 16 mixtures (compared with using only the standard feature set). A possible explanation for this improvement is that the information provided by the new (nonlinear) features deals with different aspects of the speech dynamics and therefore is valuable for the recognition process.

5 Conclusions

In this paper we have described how to apply efficient nonlinear DSP algorithms to speech signals in order to extract novel acoustic features related to their nonstationary and nonlinear dynamics of the modulation and chaotic type. Furthermore we have developed a hybrid feature set for speech recognition that includes both the standard linear features as well as the new nonlinear features and applied this new feature set to HMM-based phoneme recognition. Our experimental results have shown a significant improvement in recognition over the TIMIT database.

Given the relation of the underlying nonstationary and nonlinear models to the physics and the true dynamics of speech production and given the efficiency of the nonlinear DSP algorithms we have developed to extract the corresponding nonlinear features, we believe that the modulation and chaotic models and related nonlinear algorithms have a strong potential in speech recognition.

In the near future, we intend to apply the modulation and chaotic features for speech recognition in noisy environments and for large vocabulary speech recognition.

Regarding the modulations, the Spline-ESA can offer robustness in the speech demodulation problem. Other goals of our on-going research include: experimentation with more sophisticated

chaotic features, such as generalized dimensions and Lyapunov exponents which contain dynamical information; a better integration of chaotic features with modulation features; improvement of the algorithms for extracting chaotic features in the presence of noise.

References

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[2] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[3] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal, Eds., NATO Advanced Study Institute Series D, vol. 55, Bonas, France, July 1989.

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

[5] D. J. Tritton, *Physical Fluid Dynamics*, 2nd edition, Oxford Univ. Press, New York, 1988.

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

[7] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[8] A. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation", *J. Acoust. Soc. Amer.*, 99 (6), pp.3795–3806, June 1996.

[9] A. Potamianos and P. Maragos, "Speech Processing Applications Using an AM–FM Modulation Model", *Speech Communication*, vol.28, pp.195-209, 1999.

[10] T. F. Quatieri, C. R. Jankowski and D. A. Reynolds, "Energy Onset Times for Speaker Identification", *IEEE Signal Process. Lett.*, vol.1(11), pp.160-162, Nov. 1994.

[11] H.Tolba and D. O'Shaughnessy, "Automatic speech recognition based on cepstral coefficients and a mel-based discrete energy operator", in *Proc. ICASSP-98*, Seattle, WA, pp. 973–976, May 1998.

[12] G. Zhou, J. Hansen and J. F. Kaiser, "Linear and nonlinear speech feature analysis for stress classification", in *Proc. Int. Conf. Speech & Language Processing*, Sydney, Australia, pp. 840–843, Dec. 1998.

[13] A. Potamianos and P. Maragos, "Time-Frequency Distributions for Automatic Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol.9, pp.196-200, Mar. 2001.

[14] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, NY, 1982.

[15] P. Maragos and A. Potamianos, "Fractal Dimensions of Speech Sounds: Computation and Application to Automatic Speech Recognition", *J. Acoust. Soc. Amer.*, 105 (3), pp.1925–1932, March 1999.

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

[17] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators", in *Proc. ICASSP-91*, Toronto, Canada, pp. 421–424, May 1991.

[18] D. Dimitriadis and P. Maragos, "An Improved Energy Demodulation Algorithm Using Splines", *Proc. ICASSP-01*, Salt Lake, Utah, May 2001.

[19] M. Unser, A. Aldroubi and M. Eden, "B-Spline signal processing: Part I–Theory. Part II– Efficient design and applications" *IEEE Trans. Signal Processing*, vol. 41, pp. 821–848, Feb. 1993.

[20] A. C. Bovik, P. Maragos, and T.F. Quatieri, "AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators", *IEEE Trans. Signal Processing*, vol. 41, Dec. 1993.

[21] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, Applied Mathematical Sciences, vol.68, 1993.

[22] H.O. Peitgen, H. Jurgens and D. Saupe. *Chaos and Fractals: New Frontiers of Science*, Springer Verlag, Berlin Heidelberg, 1992.

[23] P. Grassberger and I. Procaccia, "Measuring the Strangeness of Strange Attractors", Physica 9D, pp. 189-208, 1983.

[24] S. Young, *The HTK Book*, Cambridge Research Lab: Entropics, Cambridge, England, 1995.

6 Biographies

Dimitris Dimitriadis was born in Buffalo, NY, USA, in 1976. He received the Diploma degree in electrical & computer engineering from National Technical University of Athens, Greece, in 1999, where he is currently pursuing his PhD. His research interests include nonlinear signal processing, speech modelling and recognition.

Petros Maragos

Vasilis Pitsikalis received the Diploma degree in electrical & computer engineering from National Technical University of Athens (NTUA), Greece, in 2001, where he is currently pursuing his PhD. His research interests include chaotic modelling & nonlinear operators applied in speech and doppler ultrasound signals respectively.

Alexandros Potamianos (M'92) received the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece in 1990. He received the M.S and Ph.D. degrees in Engineering Sciences from Harvard University, Cambridge, MA, USA in 1991 and 1995, respectively. From 1991 to June 1993 he was a research assistant at the Harvard Robotics Lab, Harvard University. From 1993 to 1995 he was a research assistant at the Digital Signal Processing Lab at Georgia Tech. From 1995 to 1999 he was a Senior Technical Staff Member at the Speech and Image Processing Lab, AT&T Shannon Labs, Florham Park, NJ. In February 1999, he joined the Multimedia Communications Lab at Bell Labs, Lucent Technologies, Murray Hill, NJ. He is also an adjunct Assistant Professor at the Department of Electrical Engineering of Columbia University, New York, NY. His current research interests include speech processing, analysis, synthesis and recognition, dialog and multi-modal systems, nonlinear signal processing, natural language understanding, artificial intelligence and multimodal child-computer interaction.