Available online at www.sciencedirect.com
SCIENCE DIRECT
Speech Communication 48 (2006) 819–837
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Continuous energy demodulation methods and application to speech analysis ☆

Dimitrios Dimitriadis *, Petros Maragos

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

Received 5 April 2004; received in revised form 10 March 2005; accepted 31 August 2005

Abstract

Speech resonance signals appear to contain significant amplitude and frequency modulations. An efficient demodulation approach is based on energy operators. In this paper, we develop two new robust methods for energy-based speech demodulation and compare their performance on both test and actual speech signals. The first method uses smoothing splines for discrete-to-continuous signal approximation. The second (and best) method uses time-derivatives of Gabor filters. Further, we apply the best demodulation method to explore the statistical distribution of speech modulation features and study their properties regarding applications of speech classification and recognition. Finally, we present some preliminary recognition results and underline their improvements when compared to the corresponding MFCC results.

Keywords: Nonstationary speech analysis; Energy operators; AM–FM modulations; Demodulation; Gabor filterbanks; Feature distributions; ASR; Robust features; Nonlinear speech analysis

1. Introduction

Nonlinear, time-varying signal models of the AM–FM type are nowadays receiving significant attention in nonstationary signal and speech processing schemes. The estimation of the signal instantaneous frequencies and amplitude envelopes is referred to as the 'Demodulation Problem'. Demodulating AM–FM signals, i.e., nonstationary sines that have a combined amplitude modulation (AM) and frequency modulation (FM)

$$x(t) = a(t)\cos\left(\int_0^t \omega(\tau)\,\mathrm{d}\tau\right) \tag{1}$$

has been a major research problem with many applications in communication systems, speech processing, and in general, nonstationary signal analysis. To solve it, a new approach was developed in the 1990's based on nonlinear differential operators that can track the instantaneous energy (or its derivatives) of a source producing an oscillation (Kaiser, 1983, 1990; Maragos and Potamianos, 1995). The main representative of this class of operators is the continuous-time Teager–Kaiser energy operator (TEO) $\Psi[x(t)] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$, where $\dot{x}(t) = \mathrm{d}x(t)/\mathrm{d}t$.

☆ This work was partially supported by the European NoE MUSCLE.
* Corresponding author. Tel.: +302107722964.
E-mail addresses: ddim@cs.ntua.gr, ddimitriadis@gmail.com (D. Dimitriadis), maragos@cs.ntua.gr (P. Maragos).
URL: http://cvsp.cs.ntua.gr (D. Dimitriadis).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.08.007

Applied to the AM–FM signal (1), Ψ yields the instantaneous source energy, i.e., $\Psi[x(t)] \approx a^2(t)\omega^2(t)$, where the approximation error becomes negligible (Maragos et al., 1993) if the instantaneous amplitude $a(t)$ and instantaneous frequency $\omega(t)$ do not vary too fast or too much with respect to the average value of $\omega(t)$. Thus, AM–FM demodulation can be achieved by separating the instantaneous energy into its amplitude and frequency components. Ψ is the main ingredient of the first *energy separation algorithm* (ESA)

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

developed in (Maragos et al., 1993) and used for signal and speech AM–FM demodulation.

Motivated by the strong evidence of the existence of amplitude and frequency modulations (AM–FM) in speech resonance signals, which make their amplitudes and frequencies vary instantaneously within a pitch period (Maragos et al., 1993), we propose to model each speech resonance with an AM–FM signal

$$r(t) = A \exp\left(-\int_0^t b(\tau)\,\mathrm{d}\tau\right) \cos\left(\int_0^t \omega(\tau)\,\mathrm{d}\tau\right) \tag{3}$$

and the total speech signal as a superposition of a small number of such AM–FM signals. The smooth estimation of their instantaneous frequencies $\omega(t)$ and amplitude envelopes $|a(t)|$ is of significant importance for speech analysis and feature extraction applications.

The instantaneous energy separation methodology has led to several classes of algorithms for demodulating discrete-time AM–FM signals

$$r[n] = r(nT) = A \exp\left(-\int_0^n b[k]\,\mathrm{d}k\right) \cos\left(\int_0^n \omega[k]\,\mathrm{d}k\right), \tag{4}$$

where $a[n] = a(nT)$ and $\omega[n] = T\omega(nT)$. Note that in Eq. (4) the $\int[\cdot]$ notation is used symbolically instead of the summation notation $\sum[\cdot]$ for the discrete signals involved; i.e., in the FM-part it is used to underline the differential relationship between ω and phase. A direct approach to this objective is to apply the discrete-time Teager–Kaiser operator $\Psi_d[x_n] \equiv x_n^2 - x_{n-1}x_{n+1}$, where $x_n \equiv x[n]$, to the discrete-time AM–FM signal (4) and thus derive discrete energy equations of the form

$$\Psi_d[r_n] \approx a^2[n]\sin^2(\omega[n]). \tag{5}$$

This yields the following algorithm, called *Discrete ESA* or DESA (Maragos et al., 1993)

$$\arccos\left(1 - \frac{\Psi_d[r_n - r_{n-1}] + \Psi_d[r_{n+1} - r_n]}{4\Psi_d[r_n]}\right) \approx \omega[n] = 2\pi f[n], \tag{6}$$

$$\sqrt{\frac{\Psi_d[r_n]}{\sin^2(\omega[n])}} \approx |a[n]|. \tag{7}$$

Another approach involves estimating the instantaneous frequency by modeling the discrete-time signal r_n via the exact Prony, as shown in (Fertig and McClellan, 1996; Ramalingam, 1996). This yields algorithms that also contain the discrete energy operator as their main ingredient. The advantages of the ESAs are efficiency, low computational complexity and excellent time resolution (5-sample window). Their main disadvantage, though, is a moderate sensitivity to noise (Potamianos and Maragos, 1994).

In this paper, two more systematic demodulation approaches are developed where at first, the discrete-time signal is expanded in the continuous-time domain and then, the continuous-time ESA of Eq. (2) is applied upon. The first approach is to approximate the signal with smoothing splines and then differentiate the approximating continuous-time polynomials (Dimitriadis and Maragos, 2001). The second one combines time-differentiation and filtering of the signal with a Gabor filterbank into convolutions of the signal with the time-derivatives of the Gabor filter's impulse response. The advantages of such approaches are that we can both avoid having the noisy one-sample discrete-time approximations of the derivatives and also succeed in having smoother estimates of the signal's time-derivatives in the presence of noise.

Another contribution of this paper is the application of these novel proposed algorithms for speech analysis and feature extraction. Significant conclusions have been drawn from the instantaneous signals and their distributions. Finally, features inspired from the speech resonance model, Eq. (3), are extracted. The motivation for such feature analysis scheme stems from improvements in ASR recognition rates which we have observed in previous experimental work (Dimitriadis and Maragos, 2003; Dimitriadis et al., 2002). In this paper, we study their distributions and dependencies on various aspects of the speech signal. Finally, we present some recognition results indicating the improvement of ASR rates in clean and noisy speech when the AM–FM feature vectors augment the MFCCs.

The paper is organized as follows: Section 2 describes the smoothing spline approach presenting, at the beginning, a brief background on splines. Section 3 presents the differentiated Gabor filtering approach. In Section 4 the experimental results, comparing these two new approaches with the standard DESA, are presented and the best algorithm is proposed. In Section 5, we apply this algorithm to speech analysis and feature extraction. We examine their statistical distributions and classification properties and we apply them to clean and noisy speech recognition tasks. Finally, in Section 6 some overall conclusions are presented.

2. Spline ESA

In this section we interpolate[1] discrete-time signals with splines. Spline functions are piecewise continuous polynomials assembled as linear combinations of B-splines. A spline function of order n has continuous (and smooth) derivatives up to order $n-1$, a very important property when using the Ψ-operator. This was our principal motivation for introducing the use of splines. In addition, smoothing splines provide even smoother time-derivatives than the exact splines, without losing the property of continuity.

2.1. Exact splines

Given the initial signal samples $x[n]$, where $n = 1,\ldots,N$, the interpolating spline function is given by

$$s_v(t) = \sum_{n=-\infty}^{+\infty} c[n]\beta_v(t-n), \tag{8}$$

where $\beta_v(t)$ is the B-spline of order v and the coefficients $c[n]$ depend only on the data $x[n]$ and the analytic expression of the B-spline. The B-spline of order v can be formed as the $(v+1)$th-fold convolution of the zeroth-order B-spline with itself

$$\beta_v(t) \equiv \underbrace{\beta_0(t) * \beta_0(t) * \cdots * \beta_0(t)}_{(v+1)\text{-times}},$$

where the zeroth-order B-spline is defined by

$$\beta_0(t) = \begin{cases} 1 & \text{if } |t| < 1/2, \\ 1/2 & \text{if } |t| = 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Using the discrete B-spline $b_v[n] \equiv \beta_v(n)$, Eq. (8) becomes

$$s_v(n) = (c * b_v)[n]. \tag{9}$$

For the exact interpolation problem, $s_v(n) = x[n]$. By transforming Eq. (9) in the Z domain, we obtain

$$C(z) = \frac{X(z)}{B_v(z)}. \tag{10}$$

Thus, the spline coefficients $c[n]$ can be determined recursively from the above equation. This approach can be considered, also, as a filter with frequency response $H(z) = 1/B_v(z)$ that is called *spline filter*. Spline filters of odd orders are proved to be always stable (Unser et al., 1991). Each original sample $s_v(n) = x[n]$ is resynthesized by the contribution of v neighbor spline coefficients.

2.2. Smoothing splines

We have used exact splines to improve the performance of the ESA and tested them on noisy AM–FM signals with different levels of SNR. The results were, however, disappointing as the exact fitting of the curve was the source of large estimation errors due to the presence of noise. The problem of noise led us to the need for approximating signal samples with the use of smoothing splines. The main advantage of smoothing splines is that the approximating polynomial does not pass precisely through the signal samples but "close enough".

The "smoothing spline approximating function" is defined as the function s_v of order $v = 2r-1$ that minimizes the mean square error criterion (Unser et al., 1993).

$$\epsilon = \underbrace{\sum_{n=-\infty}^{+\infty} (x[n] - s_v(n))^2}_{\epsilon_d} + \lambda \underbrace{\int_{-\infty}^{+\infty} \left(\frac{\partial^r s_v(x)}{\partial x^r}\right)^2 \mathrm{d}x}_{\epsilon_s},$$

where ϵ_d is the mean square error of the approximation function and ϵ_s is the mean square error introduced by the need for a smoothed curve. This criterion is a compromise between the need for closer-to-the-data points approximation curve and the need for a smoothed curve. The positive parameter λ quantifies the approximating curve's smoothness, criterion ϵ_s, and how close to the data points

[1] We have also experimented with least-square polynomials for the interpolation process. However, the main problem using such polynomials is that there is no theoretical guaranty that the time-derivatives of the interpolant are smooth.

this curve pass, criterion ϵ_d. For $\lambda = 0$, no smoothing effect takes place and the approximation curve fits exactly the signal samples. If $\lambda \neq 0$, the deviation from the data samples increases with λ in a nonlinear way, while concurrently the smoothing performance is increasing, too. r is the order of the polynomial's time-derivative that regulates its smoothness.

As shown in (Unser et al., 1993), the approximating polynomial $s_v(t)$ minimizing the mean square error ϵ is a linear combination of splines β_v, as in Eq. (8), though the coefficients $c[n]$ are computed as the output of an IIR filter $H_v^\lambda(z)$, much different from the filter in (10)

$$C(z) = H_v^\lambda(z)X(z) = \frac{X(z)}{P_v^\lambda(z)}, \tag{11}$$

where $P_v^\lambda(z)$ is given by

$$P_v^\lambda(z) = B_v(z) + \lambda(-z + 2 - z^{-1})^r \tag{12}$$

and r is defined as above. The IIR filter $H_v^\lambda(z)$ has a symmetric impulse response and all its poles lie inside the unit circle of the complex plane (Aldroubi et al., 1992). So, spline coefficients $c[n]$ can be stably determined via a few recursive equations (Unser et al., 1993). Henceforth, smoothing splines with $\lambda \neq 0$ are applied in every case, unless stated otherwise.

2.3. Spline ESA

The previous discussion leads us to approximate a discrete-time signal $x[n]$ using smoothing splines of vth-order and thus create a continuous-time signal

$$s_v(t) = \sum_{n=-\infty}^{+\infty} c[n]\beta_v(t-n). \tag{13}$$

The basic idea of the new approach for ESA-based demodulation is to apply the continuous-time energy operator Ψ and the continuous ESA to the continuous-time signal $s_v(t)$ instead of using the discrete energy operator Ψ_d and the DESA on the discrete signal $x[n]$

$$\Psi[s_v(t)] = \left[\frac{\partial s_v(t)}{\partial t}\right]^2 - s_v(t)\frac{\partial^2 s_v(t)}{\partial t^2}. \tag{14}$$

The use of continuous ESA requires the estimation of the signal's first-, second- and third-order derivatives. A basic property of the B-splines is that their time-derivatives can be obtained from a recursive equation

$$\frac{\mathrm{d}\beta_v(t)}{\mathrm{d}t} = \beta_{v-1}\left(t+\frac{1}{2}\right) - \beta_{v-1}\left(t-\frac{1}{2}\right). \tag{15}$$

The time derivative of the polynomial (13) is given by

$$\frac{\partial s_v(t)}{\partial t} = \sum_{n=-\infty}^{+\infty} c[n]\frac{\partial \beta_v(t-n)}{\partial t}. \tag{16}$$

Given the coefficients $c[n]$ of the spline approximation (13) and taking under consideration (15), after some math, we can derive the following closed-form expressions of these derivatives, involving only the coefficients $c[n]$ and the B-spline analytic expressions

$$\frac{\partial s_v(t)}{\partial t} = \sum_n (c[n] - c[n-1])\beta_{v-1}(t-n+1/2), \tag{17}$$

$$\frac{\partial^2 s_v(t)}{\partial t^2} = \sum_n (c[n+1] - 2c[n] + c[n-1])\beta_{v-2}(t-n), \tag{18}$$

$$\frac{\partial^3 s_v(t)}{\partial t^3} = \sum_n (c[n+1] - 3c[n] + 3c[n-1] - c[n-2]) \cdot \beta_{v-3}(t-n+1/2). \tag{19}$$

By applying these signal derivatives in the continuous ESA, we can estimate the instantaneous amplitude $a(t)$ and frequency $\omega(t)$ of the continuous signal $s_v(t)$. Finally, we obtain the sampled estimates of the instantaneous amplitude and frequency signals ($a[n] = a(nT)$, $\omega[n] = T\omega(nT)$) of the original discrete signal $x[n]$. This whole approach presented above is called the *Spline ESA*.

An important part of the Spline ESA is the computation of the spline coefficients $c[n]$. In the sequel, the details of this algorithm are discussed. First, the zeros of the denominator polynomial $P_v^\lambda(z)$ in Eq. (11) are estimated. Due to the symmetric form of this polynomial the zeros come in pairs (z_i, z_i^{-1}), $i = 1, \ldots, r$. Thus, the transfer function in Eq. (11) can be formulated as

$$H_v(z) = c_0 \prod_{i=1}^{r} \frac{-z_i}{(1 - z_i z^{-1})(1 - z_i z)}, \tag{20}$$

where c_0 is a normalizing constant depending on the spline order v. From Eqs. (11) and (20) (Unser et al., 1993) the recursive equations are:

$$\begin{aligned} &y_i^+[n] = y_{i-1}[n] + z_i y_i^+[n-1], \quad n = 2, \ldots, N, \\ &y_i[n] = z_i(y_i[n+1] - y_i^+[n]), \quad n = N-1, \ldots, 1, \\ &y_i[N] = a_i(2y_i^+[N] - y_{i-1}[N]) \end{aligned} \tag{21}$$

where $a_i = -z_i/(1 - z_i^2)$, $y_{i-1}[n]$ is the input and $y_i[n]$ is the output of a digital filter with transfer function

$$T_i(z) = \frac{-z_i}{(1 - z_i z^{-1})(1 - z_i z)}$$

and $y_0[n] = x[n]$. If the above step is repeated as many times as the number of pole pairs (z_i, z_i^{-1}), the final output sequence $y_r[n]$ equals $c[n]$. The boundary conditions can be set as

$$y_i^+[1] = \sum_{k=1}^{k_0} z_i^{|k-1|} y_{i-1}[k],$$

where k_0 is an integer that ensures a certain level of precision, defined a priori.

An example is presented for two different values of λ, $\lambda = 0$ and $\lambda = 0.5$ to clarify the estimation process of $c[n]$ for splines of order $v = 5$ ($r = 3$). First, when $\lambda = 0$, we interpolate the input signal using exact splines of order $v = 5$. The denominator of the transfer function $H_5(z)$ is $P_5(z) \equiv B_5(z)$ and the poles are

$$z_1 = -0.04309, \quad z_2 = 0.43057,$$
$$z_3 = z_1^{-1}, \quad z_4 = z_2^{-1}.$$

Setting $\lambda = 0.5$ (when $v = 5$) in Eq. (12) the poles of $H_5(z) = 1/P_5(z)$ are

$$z_1 = 0.32548,$$
$$z_2 = 0.32154 - 0.47128\mathrm{i}, \quad z_3 = 0.32154 + 0.47128\mathrm{i},$$
$$z_4 = z_1^{-1}, \quad z_5 = z_2^{-1}, \quad z_6 = z_3^{-1}.$$

In both cases, we find $c[n]$ using the algorithm of Eq. (21), even though the number and values of the poles are quite different. Having computed $c[n]$, the coefficient sequence is convolved with the B-spline b_v to yield the approximating signal s_v. In general, the evaluation of the spline coefficients by the filtering approach, presented above, is less computationally intensive than the standard numerical analysis approach using sparse Toeplitz matrices.

The choice of spline order of $v = 5$ is the result of both experimentation and theoretical analysis. The use of the Spline ESA requires the estimation of the signal (or its continuous-time expansion) time-derivatives up to those of third-order. It becomes obvious that the smallest possible order of B-spline with continuous derivatives is $v = 5$. Besides the theoretical need for continuity of the derivatives (for the use of CT-ESA), several experiments were undertaken for different B-spline order and the best

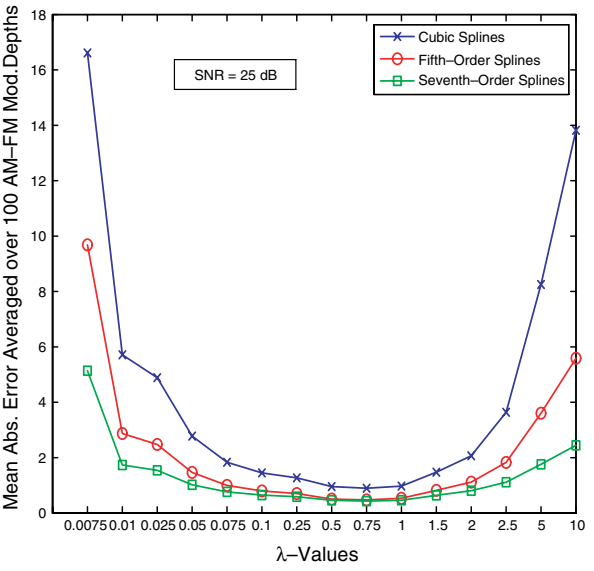

Fig. 1. Mean absolute frequency estimation error of spline ESA as a function of the smoothing parameter λ when SNR = 25 dB.

results were again obtained for the fifth-order splines, as presented in Fig. 1. Finally, only five samples are needed for the estimation of the spline coefficients $c[n]$, when selecting $v = 5$ and the time-resolution remains the same as the corresponding one of the DESA. For larger values of v the time-resolution becomes successively poorer.

The choice of the optimal value of λ is not completely arbitrary, too. We have attempted to experimentally determine a suitable range of values for the λ-parameter for different SNRs. In Fig. 1, the mean absolute errors of the cubic, fifth- and seventh-order smoothing spline polynomials are presented for various values of the λ parameter when SNR = 25 dB. For a given AM–FM signal, Eq. (25), Spline ESA is applied to estimate the demodulating error performance as a function of the λ-parameter and the spline order. In these experiments the corresponding mean absolute curves show global minima for particular range of λ-values around 0.75. More specifically, the global minima occur when $\lambda \in [0.1, 1]$ independently of the SNR values. The mean absolute errors of the seventh-order smoothing spline approximation are always smaller than the corresponding ones of the cubic and fifth-order spline polynomials. Noticing that, the global minimum values are quite similar for these two approximation curves (the fifth-and seventh-order spline polynomials), we are proposing the use of the fifth-order smoothing spline approximation

polynomial since it yields almost the optimal minimum error rates and has better time-resolution, five sample window instead of seven. The optimal value of λ is not a priori known and can be determined only through experimentation because the estimation errors implicitly depend on the SNR, the signal, and the application.

3. Gabor ESA

The ESA cannot handle wideband signals, such as speech signals, due to inherent limitations of the algorithm. One efficient way to deal with such limitations is bandpass filtering of the signal. For this process, the Gabor filters are chosen for several reasons listed in (Maragos et al., 1993), such as the optimal time-frequency discriminability. In Section 2, smoothing splines are used to approximate the discrete-time signals $x[n]$ and then they are derived using closed formulas. The continuous-time TEO Ψ, combined with bandpass filtering and sampled at time instances $t = nT$, is given by

$$\Psi[s(t)] = \dot{s}^2(t) - s(t)\ddot{s}(t)|_{t=nT}, \quad s(t) = x(t) * g(t), \tag{22}$$

where $x(t)$ is the continuous-time signal and $g(t)$ is the Gabor filter impulse response

$$g(t) = \exp(-\beta^2 t^2)\cos(\omega_c t). \tag{23}$$

The constants β and ω_c are the filter parameters.

Since convolution commutes with time-differentiation (Papoulis, 1962)

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m}(x(t) * g(t)) = x(t) * \frac{\mathrm{d}^m}{\mathrm{d}t^m} g(t), \quad m = 1, 2, 3, \ldots \tag{24}$$

The Gabor time-derivatives are given by closed formulae

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = \left(-2\beta^2 t\cos(\omega_c t) - \omega_c \sin(\omega_c t)\right)\exp(-\beta^2 t^2),$$

$$\frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2} = \left(4\beta^2\omega_c t\sin(\omega_c t) + (4\beta^4 t^2 - 2\beta^2 - \omega_c^2)\cos(\omega_c t)\right) \times \exp(-\beta^2 t^2),$$

$$\frac{\mathrm{d}^3 g(t)}{\mathrm{d}t^3} = \left[\left(12\beta^4 t - 8\beta^6 t^3 + 6\beta^2\omega_c^2 t\right)\cos(\omega_c t) + \left(6\beta^2\omega_c - 12\beta^4\omega_c t^2 + \omega_c^3\right)\sin(\omega_c t)\right]\exp(-\beta^2 t^2).$$

Using the above equations in Eq. (22), the output of Ψ acting on the bandpass filtered signal is given by

$$\begin{aligned}\Psi[s(t)] &= \Psi[x(t) * g(t)] \\ &= \left[\frac{\mathrm{d}}{\mathrm{d}t}(x(t) * g(t))\right]^2 - (x(t) * g(t))\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}(x(t) * g(t))\right] \\ &= \left[x(t) * \frac{\mathrm{d}g(t)}{\mathrm{d}t}\right]^2 - (x(t) * g(t))\left[x(t) * \frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2}\right].\end{aligned}$$

Through this approach, the necessary processes of bandpass filtering and the subsequent differentiations are combined into a single convolution with derivatives of the Gabor filter's impulse response.

Since the output of the continuous-time TEO Ψ will be sampled at time instances $t = nT$, to implement the Gabor TEO we must essentially convolve the discrete-time speech signal $x[n]$ with the sampled derivatives of the Gabor function

$$g^{(m)}[n] = \frac{\mathrm{d}^m}{\mathrm{d}t^m} g(t)|_{t=nT}.$$

The next step is to incorporate the Gabor TEO into the continuous-time ESA formulae. The resulting demodulation algorithm is called the *Gabor ESA*. This algorithm exhibits some advantages compared to the Spline ESA or to the original discrete demodulation algorithm DESA. First, bandpass filtering of noisy signals increases the SNR of the filtered signals. Second, fewer parameters are required, compared to the Spline ESA where the λ-parameter is important. The parameters needed are only those concerning the filterbank specifications. Finally, the differentiation is introduced on the filters and not on the speech signal itself and this fact leads to smoother results.

The Gabor ESA is computationally more intensive than the original DESA or the Spline ESA when they are applied upon bandpass filtered signals. On the other hand, as shown in Fig. 2, Gabor ESA provides smoother estimates of the instantaneous frequency compared to the corresponding ones of the DESA, especially on noisy signals.

In Fig. 2, a phoneme */aa/* extracted from the TIMIT database is filtered by a Gabor filter with center frequency $f_c = 1285$ Hz and bandwidth $\beta = 400$ Hz. The filter is manually placed according to the phoneme's energy spectrum to filter just one resonance. White Gaussian noise with SNR = 10 dB is added to examine the algorithm robustness to noise. In general, both Gabor ESA and DESA provide robustness to noise, but the Gabor ESA yields somewhat smoother estimates, as shown in

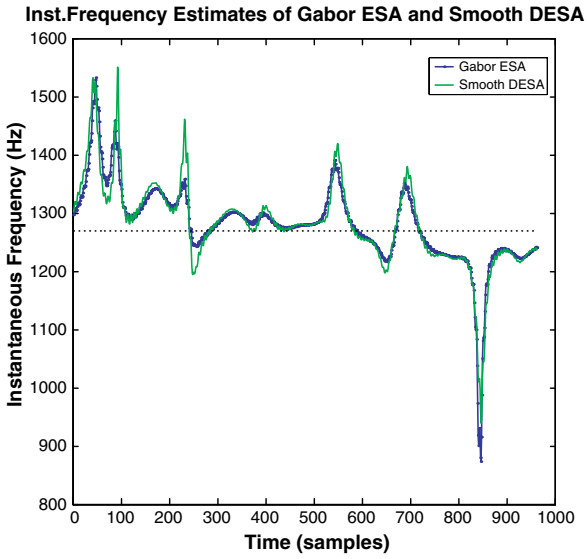

Fig. 2. Instantaneous frequency signal of a bandpass filtered phoneme /aa/using Gabor ESA and DESA (dotted line indicates the filter's center frequency).

Fig. 2. This is due to the fact that the differentiation is held on the filters and not on the signal. Thus, the differentiation does not introduce additional noise such as the one introduced by the DESA's approximation of the signal derivatives.

4. Experimental results

4.1. AM–FM test signals

The first series of experiments are performed with known input AM–FM signals. The test signals have different AM and FM modulation depths and are the same as those used in (Maragos et al., 1993). This family of AM–FM test signals is

$$s[n] = \left(1 + k\cos\left(\frac{\pi n}{100}\right)\right)\cos\left[\frac{\pi n}{5} + \frac{\ell}{\pi/100}\sin\left(\frac{\pi n}{100}\right)\right] + W[n], \tag{25}$$

where $n = 1,\ldots,400$, and $(k,\ell) = (0.05i, 0.05j)$, $i,j = 1,\ldots,10$ are, respectively, the AM and FM modulation depths (indexes). Also, zero-mean, gaussian noise $W[n]$ of different SNR levels is added to examine the algorithm robustness to noise. The estimation errors, for different SNR levels, are averaged over 100 different AM–FM depths. The process followed in this experiment is the same for all three algorithms. First, the input AM–FM signal for a given SNR level and certain AM and FM modulation indexes is estimated and then, bandpass filtering is introduced. The Gabor ESA is based on such filtering process so, to keep the comparison of the corresponding error rates fair (input signals should have the same SNR), the filtering process is held for the DESA and the Spline ESA, too. This process increases the SNR of the filtered signals and consequently improves the algorithm robustness to noise. The Gabor filter is placed at the mean instantaneous frequency value $\Omega_c = \pi/5$ and its bandwidth parameter is $\beta = \pi/12$, so that the signal falls within the filter's passband. Then, the demodulated instantaneous signals, $a[n]$ and $f[n]$, are estimated and compared to the reference instantaneous amplitude and frequency signals

$$a_k[n] = 1 + k\cos\left(\frac{\pi}{100}n\right) \quad \text{and}$$

$$f_\ell[n] = \frac{1}{2\pi}\left(\frac{\pi}{5} + \ell\cos\left(\frac{\pi}{100}n\right)\right).$$

The gain of the instantaneous amplitude signal $a_k[n]$ is altered due to the filtering process and a filter compensation algorithm is followed as proposed in (Bovik et al., 1993).

As shown in Fig. 3, the instantaneous frequency estimation error rates are quite small and the filtering process adds additional robustness to the demodulation algorithms since the error rates seem to be invariant to the SNR levels.

Spline ESA's performance is strongly dependent on the λ-parameter. However, to the best of our knowledge, there is no efficient way to optimally adjust this parameter for minimizing the demodulation error. Thus, the use of the Gabor ESA is proposed for the demodulation scheme, since it is simpler and yields smaller errors than both the Spline ESA and the DESA. This choice is, also, supported by the next series of experiments with speech signals, where the Gabor ESA clearly outperforms the other two algorithms.

4.2. Speech test signals

In this section, the proposed demodulation algorithms are applied upon speech signals. Many experiments have been conducted with different phoneme classes and the results, in terms of error rates, are proved to be quite similar. Thus, similar results can be reproduced for different kinds of phonemes, without affecting the general conclusions, although we herein present figures only for a few phonemes. All phonemes are extracted from the TIMIT speech database. Their sampling frequency

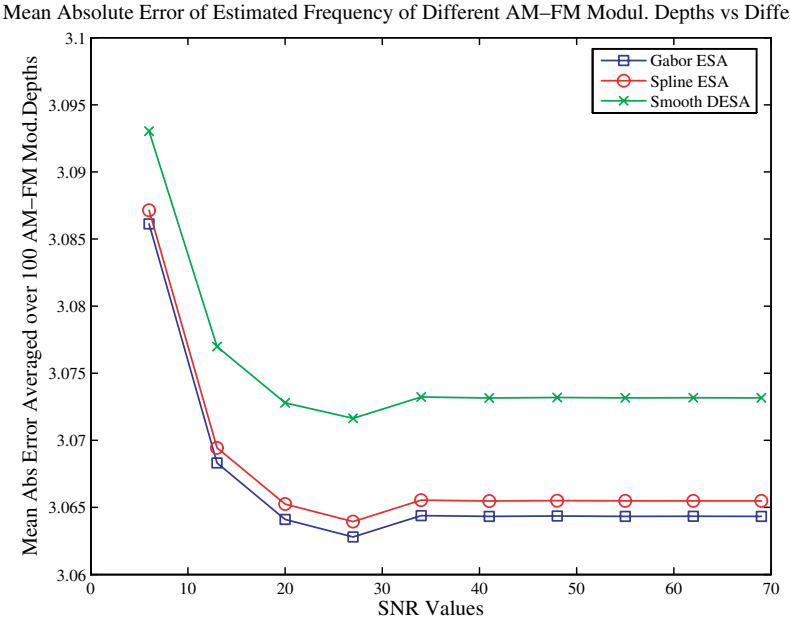

Fig. 3. Comparison of different ESAs to instantaneous frequency signals.

is $F_s = 16$ kHz. The phoneme segmentation is based on the given transcription files of the database.

There is no known method to estimate precisely and uniquely the instantaneous modulating signals for a speech signal, in general. As a consequence, it is impossible to compare the estimation errors in a quantitative way. Their performance could be either examined in a qualitative way e.g. in terms of smoothness and the existence of spikes or singularities, or compared to some reference signals, as in Fig. 7.

At first, a Gabor filter is placed manually according to the phoneme spectrum to demodulate only one of its formants. The demodulation estimates are examined in terms of their smoothness (whether or not spikes appear in the instantaneous modulating signals) and their numerical singularities, Fig. 2. Then, in a second experiment, a Gabor filterbank is used to concurrently test the algorithm performance using various filters. The filterbank is mel-spaced spanning the interval $[0 - F_s/2]$ Hz and six filters, in total, are used with a frequency overlap of 50%, Table 1. In Fig. 4, the afforementioned filterbank is superimposed over the spectrum of a phoneme */aa/*. The filterbank parameters are the same as those used for ASR experiments. The bandpass speech outputs are demodulated using all three algorithms and their corresponding estimates are presented in Fig. 5. The continuous-time algorithms yield quite similar instantaneous estimates, thus, for the next experiment where a phoneme */aa/* is demodulated, only one of these algorithms is used. In Fig. 6, the output of the fifth-filter for a phoneme */aa/* is demodulated and the instantaneous frequency estimates are presented. The continuous algorithm provides smoother estimates, with smaller fluctuations concerning the instantaneous frequency signals.

Table 1
Mel-spaced Gabor filterbank parameters

Index	Gabor filterbank parameters	
	Center frequency (in Hz)	Bandwidth (in Hz)
1	303	606
2	738	870
3	1361	1246
4	2254	1786
5	3535	2561
6	5370	3670

The DESA is, always, outperformed by both the Gabor ESA and the Spline ESA (when the λ parameter is chosen properly) independently of the input speech signal. Parameter λ regulates the Spline filter bandwidth, Eqs. (11) and (12), where smoother approximating curves correspond to filtering out the higher frequency part of the signal spectrum. This Spline filter is a lowpass filter and its passband is regulated by the λ parameter. Choosing too large a value

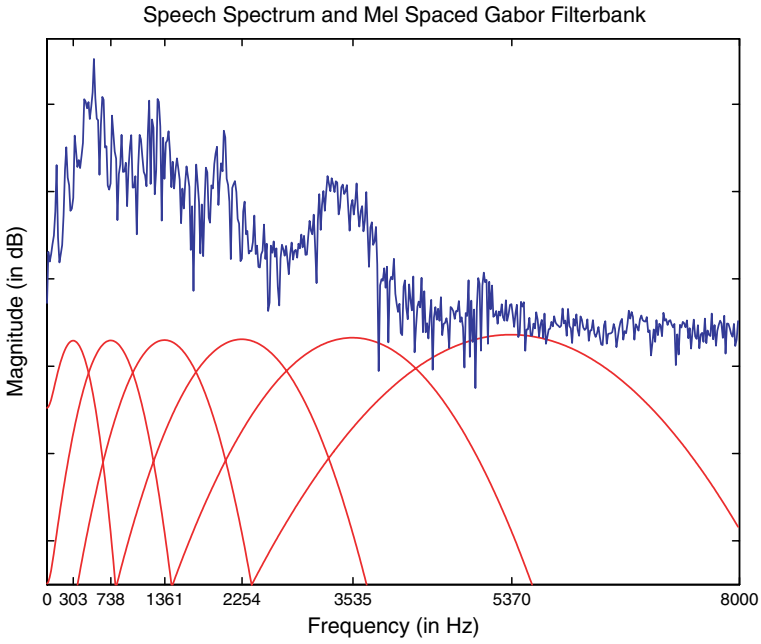

Fig. 4. Gabor filterbank superimposed over speech spectrum of a phoneme /aa/.

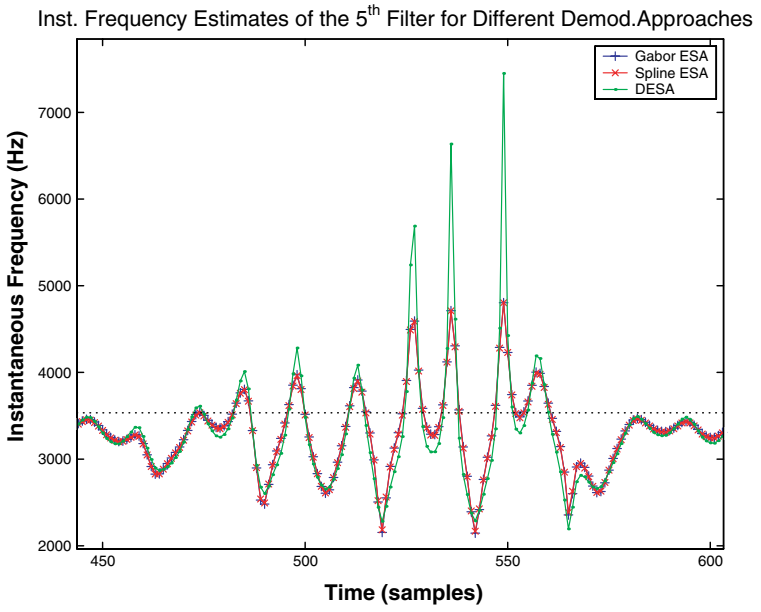

Fig. 5. Comparison of different ESAs to instantaneous frequency estimates of bandpass filtered speech signal (dotted line indicates the filter's center frequency).

for λ causes problems of oversmoothing, eliminating a significant part of the input signals' spectrum, i.e., loosing some part of the modulation signals.

Finally, one more set of experiments is conducted to examine the performance of the demodulation algorithms in different parts of the speech spectrum, using a Gabor filterbank of varying parameters. The corresponding instantaneous signals of speech cannot be determined exactly. So, the clean bandpass speech signal estimates $a_i^0[n]$ and $f_i^0[n]$ (where $i = 1, \ldots, N$ and N the number of filters) are used as reference signals. White, gaussian noise is added

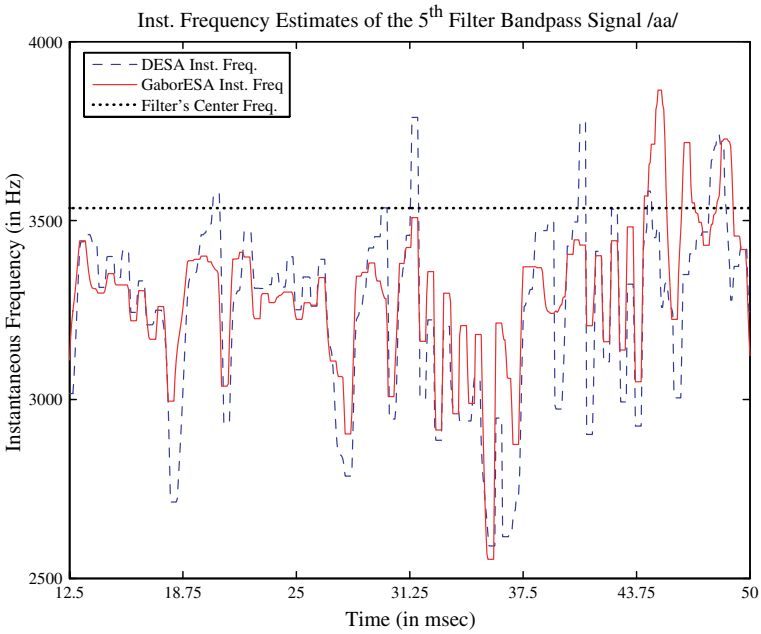

Fig. 6. Comparison of Gabor ESA and DESA to instantaneous frequency estimates of a bandpass filtered phoneme /aa/ (dotted line indicates the filter's center frequency).

to the speech signals and the yielded noisy estimates $a_i^j[n]$ and $f_i^j[n]$ (j is the index of the SNR values—here $j = 1, \ldots, 10$) are compared to the reference signals $a_i^0[n]$ and $f_i^0[n]$. The *RMS* errors are averaged over all the different noise levels. A linearly spaced Gabor filterbank, with 25 ($N = 25$) filters and filter overlap equal to 50%, is used. This experiment provides useful insight on the algorithm performance as a function of the frequency bins (indexes of the filter center frequencies) and different SNR values. The results appear to be quite similar for different phoneme classes. For this reason only one experiment, for the phoneme */aa/*, is herein presented. In Fig. 7, Gabor ESA seems to be more robust across different filter indexes and exhibits an almost flat error-rate curve, independent of the filter center frequencies. On the other hand, both DESA and Spline ESA exhibit a good behaviour for the lower part of the spectrum but they become unstable for the higher part of the spectrum, as their estimation errors increase significantly in direct proportion to the filter indexes.

The *Gabor ESA* is proven to be the best demodulating algorithm among all three candidates. The experimental results show that it is more robust in terms of noise, exhibits a uniform performance when used in different parts of the speech spectrum and finally, it yields the smoother instantaneous estimates. Also, it requires fewer parameters than the Spline ESA does. Henceforth, the Gabor ESA will be uniformly used, unless otherwise stated.

5. Distributions of speech modulation signals

The AM–FM model of speech is especially important as it provides useful insight for the formant structure of different phoneme classes, like the vowels, in time scales that the linear source-filter model considers it fixed (Rabiner and Schafer, 1978). In (Potamianos and Maragos, 1996, 1999) some first preliminary results were presented pointing that the instantaneous modulation signals appear to have different patterns depending on the phoneme classes and the speaker articulation. Herein, these dependencies are being investigated in a more detailed and thorough manner. Different phoneme classes have quite different formant structure; therefore, the demodulated instantaneous estimates $a_i[n]$ and $f_i[n]$ (where i is the filter index number) of these classes should have different distributions too. As a consequence, nonlinear speech features, based on these instantaneous signals, exhibit different distributions, and therefore, they can be useful for recognition tasks. Herein, the histograms of both the raw instantaneous signals $|a_i[n]|$ and $f_i[n]$ and the novel AM–FM features are presented for different phoneme classes. These histograms reveal

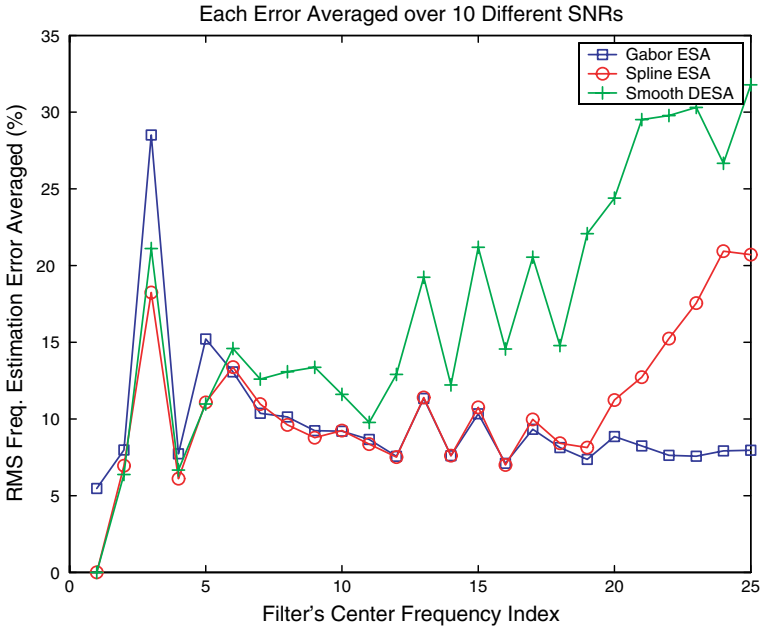

Fig. 7. Comparison of the instantaneous frequency estimates of the three demodulation algorithms on noisy versions of a phoneme /aa/ to the clean case, across A filterbank.

some part of the AM–FM formant structure of speech, potentially leading to successful pattern classification and ASR applications.

The phoneme signals are extracted from the TIMIT database according to the given transcriptions. The phoneme steady-state part is extracted (middle one third) and the proposed features are estimated over the whole span of this segment (i.e., the steady-state of the phoneme is assumed as one large speech frame). In this case, some of the transient phenomena present in speech are not taken under consideration and they are smoothed out. For the demodulation process, a mel-spaced Gabor filterbank is used as in Section 4.2. The histograms of the raw instantaneous signals are directly estimated from their corresponding values without any further postprocessing. The histogram bins span between $[0\ldots1]$ and $[0\ldots8]$ kHz for the normalized instantaneous amplitude and frequency signals, respectively. For the modulation features case, the histograms span their corresponding dynamic range. The distributions for both cases are estimated as histograms of 30 fixed, linearly spaced bins.

5.1. Distributions of instantaneous modulating signals

In this section, the raw instantaneous modulating signal distributions are being examined. First, different histograms for the same class of phonemes depending on the speaker's gender are estimated to examine whether or not the gender is an important factor affecting the structure of the instantaneous estimates. This experiment is held for several different phoneme classes and the estimated histograms for both genders appear to be quite similar. According to these experimental results, it seems that the speaker gender is not affecting the raw instantaneous signal distributions.

Further, the distributions of different phoneme classes are estimated to examine whether or not they exhibit significant differences. The vertical axis of the frequency-related histograms is plotted on the log-scale to examine whether the instantaneous frequency estimates exhibit exponential type distributions, e.g., Laplacian or Gaussian. The speech signals are randomly chosen from the TIMIT database to obtain unbiased distributions, independent of the speaker's accent and gender. Their histograms are superimposed investigating the existence of patterns in their distributions.

In Figs. 8 and 9 the distributions of the instantaneous signals $|a_i[n]|$ and $f_i[n]$ (for all six filters) are plotted for five randomly chosen instances of two different phoneme classes (phonemes */aa/* and */sh/*). The log-frequency distributions are quite linear, which leads to the conjecture that they follow an exponential distribution. Their peak is close to the

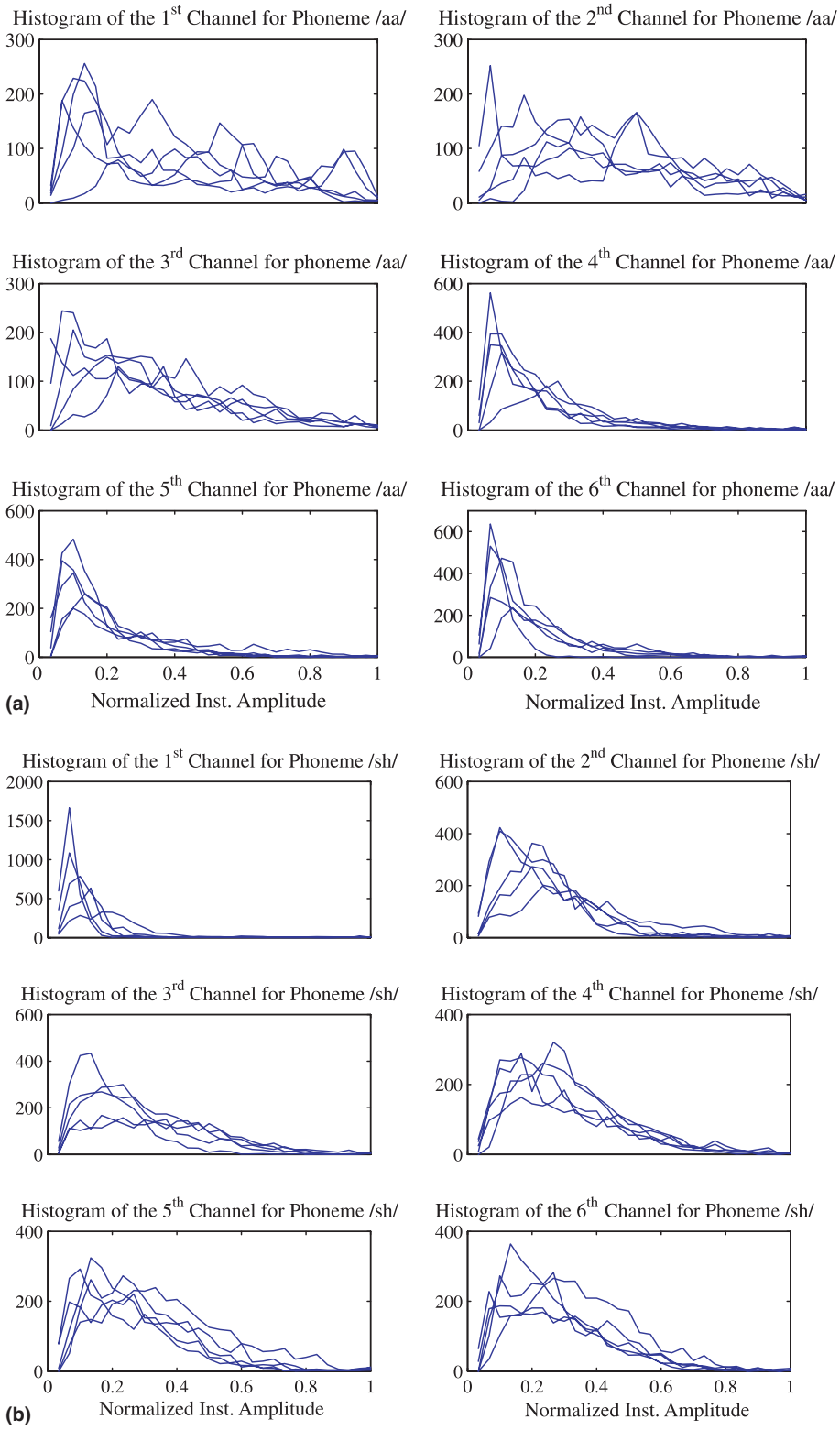

Fig. 8. (a) Five histograms of the instantaneous amplitude signals of phoneme /aa/, (b) five histograms of the instantaneous amplitude signals of phoneme /sh/.

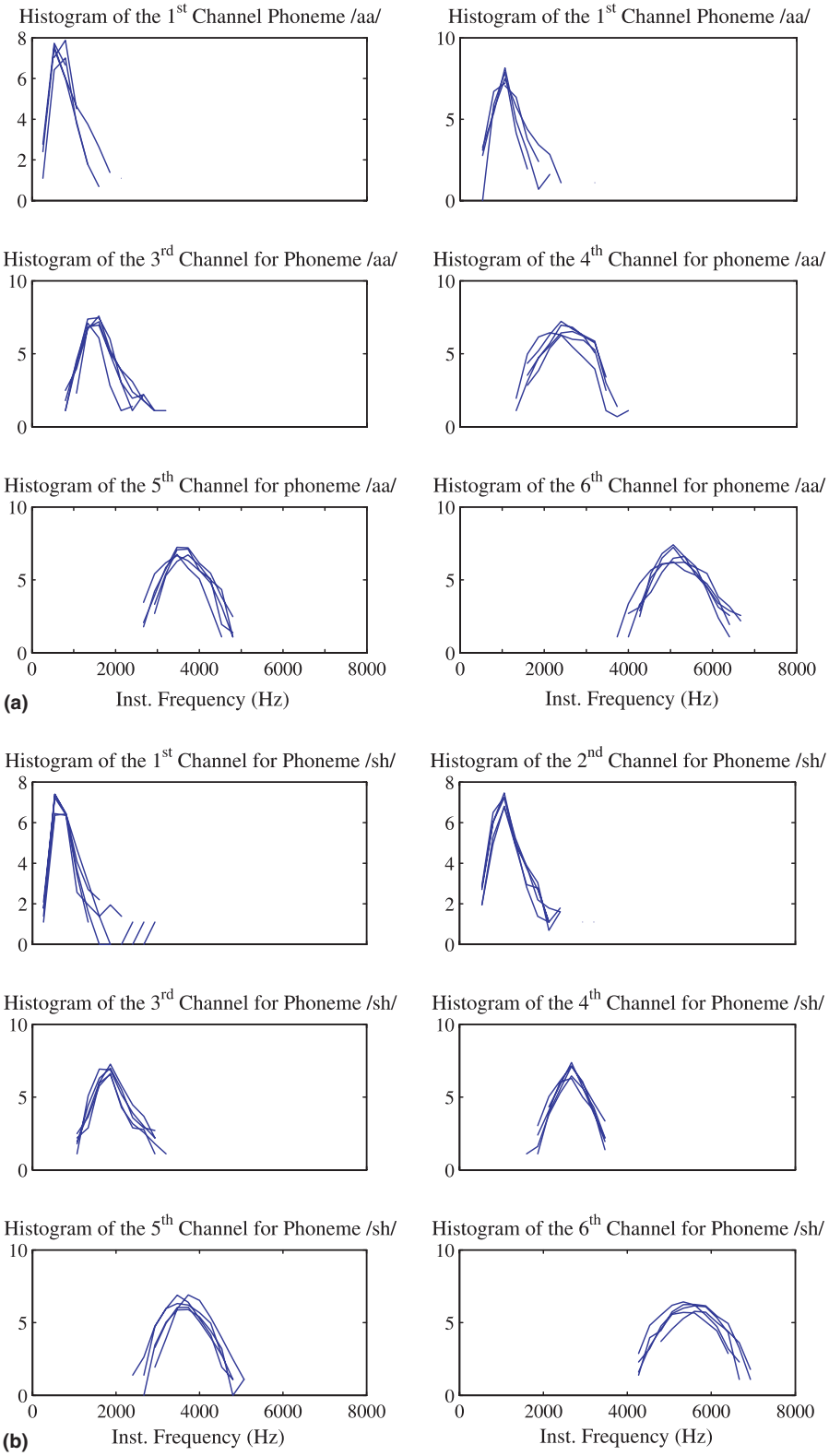

Fig. 9. (a) Five histograms of the instantaneous frequency signals of phoneme /aa/, (b) five histograms of the instantaneous frequency signals of phoneme /sh/.

center frequency f_c of the filters as dictated by the AM–FM resonance model and shown in Fig. 2. Also, the $|a_i[n]|$-distributions appear to have different patterns. For the phoneme /aa/ case, the histograms are concentrated in the lower-index bins and they exhibit small variance. On the other hand, for the phoneme /sh/ case, the distributions are presenting larger variance and span a bigger number of bins.

5.2. Distributions of modulation-related speech features

Besides the raw instantaneous signal distributions, the distributions of the proposed nonlinear modulation features are examined. These novel features consist of either *the Mean Values* of the instantaneous amplitude (*IA-Mean*), frequency (*IF-Mean*) and AM-modulating signals $b[n]$ (*BandW-Mean*) (4) or the *FM-modulation percentages* (FMP), estimated over the bandpassed speech signals, as above.

The IA-Mean and IF-Mean features are the short-time means of the normalized instantaneous amplitude and frequency signals $a_i[n]$, $f_i[n]$. The instantaneous frequency signal values have been trimmed within $\pm 5\%$ of their mean value so that isolated spikes are ignored. The FMP features are defined as $\mathrm{FMP}_i = B_i/F_i$ for each filter i, where B_i is the mean bandwidth (a weighted version of the $f_i[n]$-signal deviation (Potamianos and Maragos, 1996)) and F_i is the (amplitude) weighted mean frequency value. F_i and B_i are estimated from the information signals $a_i[n]$ and $f_i[n]$ as follows:

$$F_i = \frac{\sum_{k=0}^{T} f_i[k] a_i^2[k]}{\sum_{k=0}^{T} a_i^2[k]},$$

$$B_i = \frac{\sum_{k=0}^{T} [\dot{a}_i^2[k] + (f_i[k] - F_i)^2 a_i^2[k]]}{\sum_{k=0}^{T} a_i^2[k]}, \tag{26}$$

where $i = 1, \ldots, 6$ is the filter index and T the time window length. Finally, the AM-modulating signal $b(t)$ is defined as the ratio of the first time-derivative of the instantaneous amplitude over that signal

$$b_i(t) = -\dot{a}_i(t)/a_i(t) \tag{27}$$

and $b[n] = b(nT)$ is the sampled version of this signal (4). The proposed features provide information about the speech formant fine-structure taking advantage of the excellent time-resolution of the ESA. Thus, some of the transitional phenomena and the instantaneous formant variations are mapped onto these nonlinear AM–FM features.

The features are estimated over 1000 instances of different phonemes and their histograms are obtained. This experiment reveals different patterns for different phoneme classes. In Figs. 10–13 the distributions of the IA-Mean, IF-Mean, BandW-Mean and FMP features for the phoneme classes /aa/, /ae/, /sh/, /fl/, /p/ and /b/ are presented. In those figures, the vowel distributions seem to be different from the corresponding ones of the fricatives and the plosives. Each of the three different classes (vowels, fricatives and plosives) show similar distributions with small intra-distribution variance but on the contrary, they exhibit significant inter-distribution variance. This, however, is expected since the selected sound-classes are characterized by different formant (resonance) structure. These differences are mapped into the modulated signals. The fricative, plosive and vowel signals exhibit a very different amplitude and frequency structure as their instantaneous signals clearly indicate. Moreover, the instantaneous frequency mean values are clustered around the filter center frequencies, Fig. 11.

Figs. 10–13 indicate that there are observable differences among the collective distributions of the features for different phoneme classes. Another conclusion is that different phonemes of the same class have similar distributions over some filters, while they differentiate over the rest of their filter distributions. The classification information of the phonemes must be considered collectively over the whole filterbank of instantaneous signal. Vowels appear to have distributions with mean values well-centered in the middle bins and large variance. Fricative distributions show similar mean values but smaller variance. Finally, the distributions of plossives have small mean values but they exhibit large variance and span several bins.

The next step is to estimate the statistical moments of the distributions and to examine whether they differ sufficiently, depending on the phoneme classes analyzed. Towards this approach the following clustering experiments are applied.

5.3. PCA Plots of modulation-related speech features

Using a feature reduction technique, the feature sets are projected onto a 3D-space. The feature dimensionality is thus reduced to one half of the original set using the DCT. Thus, only the three most

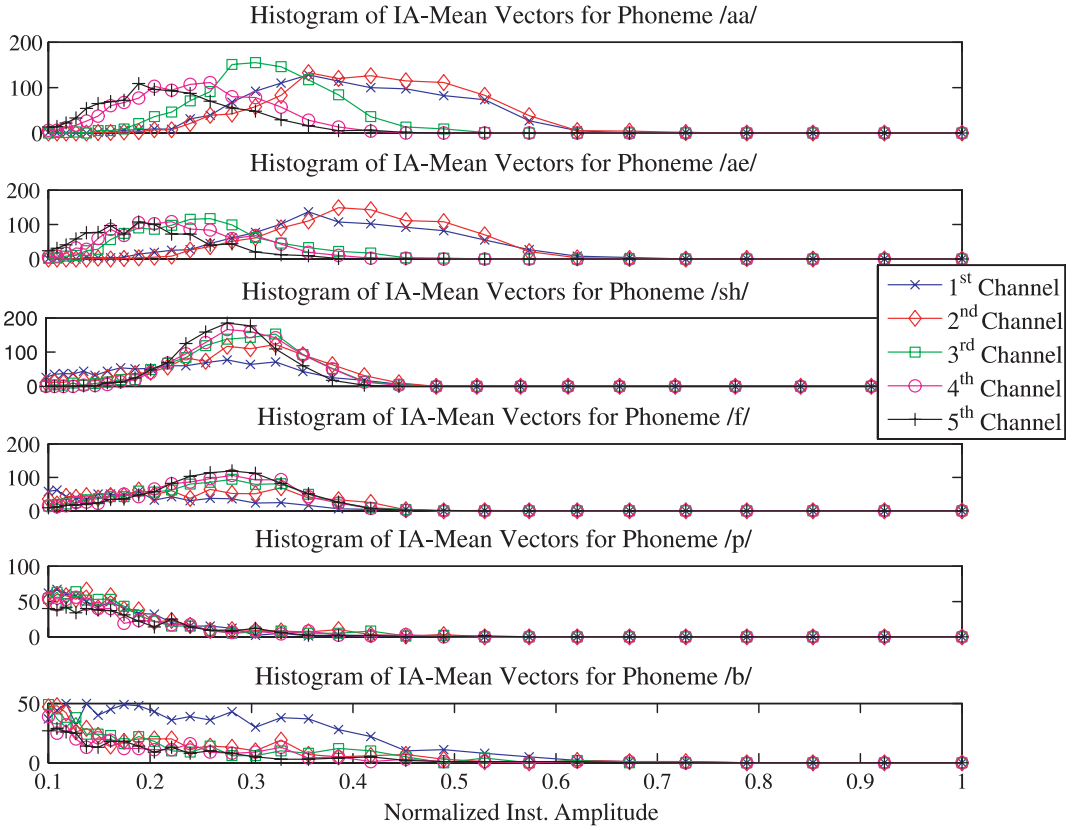

Fig. 10. Histograms of IA-Mean features for 1000 instances of phonemes /aa/, /ae/, /sh/, /f/, /p/ and /b/.

important coefficients of each feature vector are kept. This reduction is used to study the data clustering properties in a computationally more efficient way. After this process, we observe that the smoothed features (only the steady-state part of the phonemes is taken under consideration) form clusters with small overlap, as shown in Fig. 14. The 3D ellipsoids depicted, corresponding to the gaussian dispersion characteristics of the respective feature sets, are estimated by

$$(\boldsymbol{x}-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu)=const, \tag{28}$$

where μ, $\boldsymbol{\Sigma}$ are the mean vectors and the covariance matrices of the individual phoneme classes presented and *const* is the probability level corresponding to the percentage of samples included in each volume (Duda et al., 2001). Here, the ellipsoids are estimated by setting their principal axes length equal to two times the corresponding standard deviation; i.e., their x-axis length is set equal to $2\sigma_x$. In Fig. 14, the 3D-projections of the proposed nonlinear feature distributions, when examined in pairs, exhibit some separability and thus, these features could be useful in speech recognition tasks by following some well established pattern classification methodology.

5.4. Speech recognition using modulation-based features

The proposed features are applied to clean and noisy recognition tasks, after studying their distributions and classification properties. The goal is to examine whether the correct phoneme recognition rates can be improved when the standard MFCC feature set is augmented by the proposed nonlinear features and if they show additional robustness to noise. The speech databases used were obtained from the TIMIT database after adding two different kinds of noise and fixing their SNR level equal to 10 dB. Specifically, we created these *TIMIT + Noise* databases by adding white and pink noise only to the test set of the original TIMIT database, leaving the training set unaltered. We have used the HTK Toolkit (Young et al., 2002) as the HMM-based recognizer. The HMMs are 3-state,

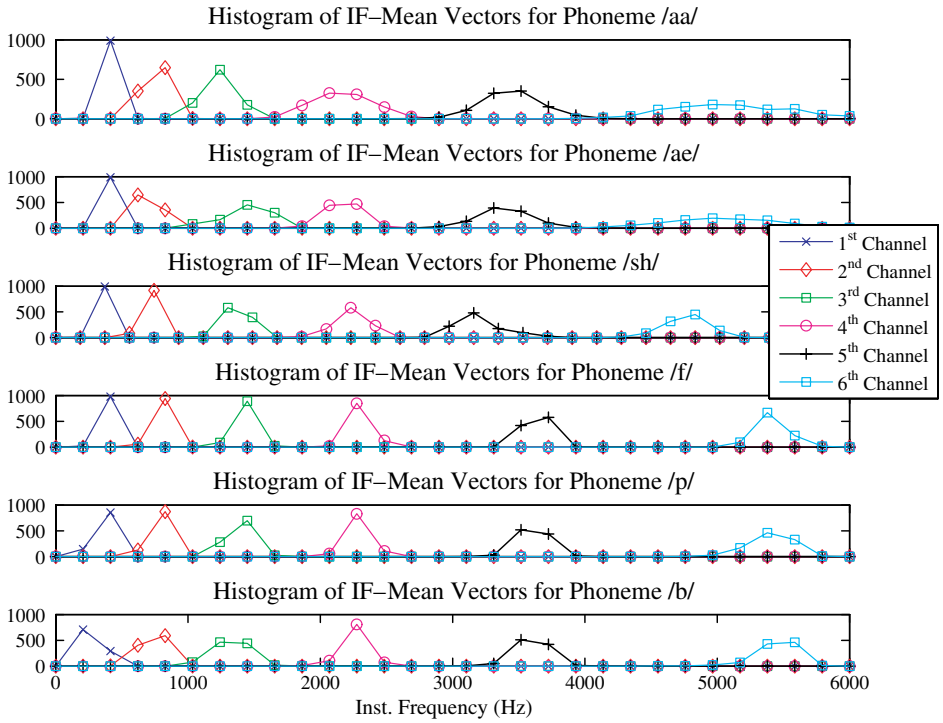

Fig. 11. Histograms of IF-Mean Features for 1000 instances of phonemes /aa/, /ae/, /sh/, /f/, /p/ and /b/.

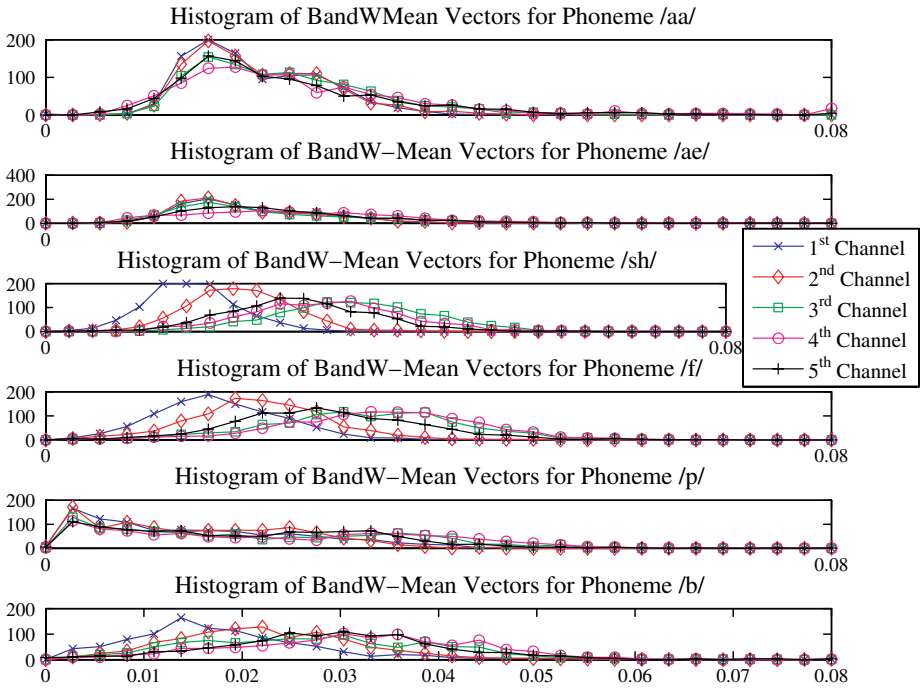

Fig. 12. Histograms of BandW-Mean features for 1000 instances of phonemes /aa/, /ae/, /sh/, /f/, /p/ and /b/.

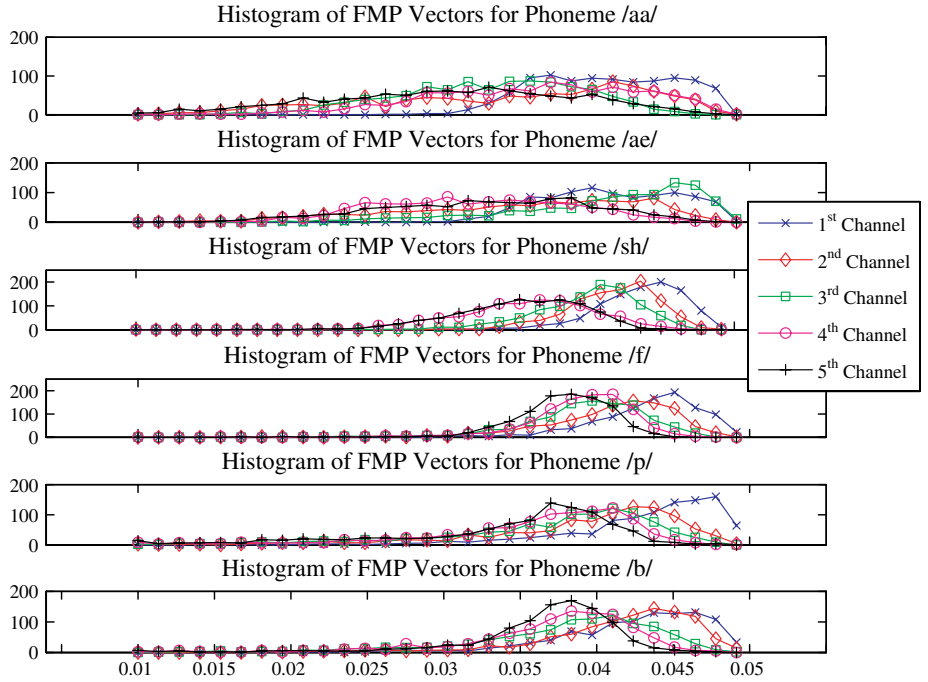

Fig. 13. Histograms of FMP Features for 1000 instances of phonemes /aa/, /ae/, /sh/, /f/, /p/ and /b/.

left–right with 16 gaussian mixtures per state. The grammar used for all cases is the all-pair, unweighted grammar. All HMM models are trained in the clean speech training set and tested in the noise-corrupted versions of the testing set.

The input vectors are split into two different data streams, one for the standard features MFCC and the other for the modulation-based features. The data streams are assumed independent (Young et al., 2002). The augmented feature vector consists of 57 coefficients, 39 samples for the 'standard' features (normalized energy, MFCCs, first and second time-derivatives) and 18 for the modulation features (6 coefficients plus their first and second time-derivatives). The frame length is set equal to 30 ms with frame-period equal to 10 ms. The weights of these two independent data streams are set $s_1 = 1.00$ and $s_2 = 0.50$, for the MFCCs and the modulation features, correspondingly. In Table 2 the recognition results are presented for the TIMIT-based ASR tasks. By combining MFCCs with the proposed AM–FM features, a performance improvement is obtained for the clean and especially for the noisy tasks where the improvement is larger (*TIMIT + Noise*) driving us to the conclusion that the proposed nonlinear features exhibit additional robustness to noise. The mean relative improvement is ranging from 18% to 25% and is achieved when the MFCC feature vectors are augmented with the proposed nonlinear modulation features.

6. Conclusions and discussion

In this paper, two new continuous-time methods are proposed for the standard speech demodulation approach (ESA). It is shown that these methods exhibit better demodulation performance, especially when noisy input signals are concerned. The best of these algorithms (the Gabor ESA) is then employed for modulation speech analysis and feature extraction.

Several useful conclusions are deduced from the proposed speech analysis scheme. First, our experiments have shown that the instantaneous amplitude and frequency signals do not seem to depend on the speaker genders since their individual distributions seem to be quite similar. These signals seem to be dependent only on the phoneme classes, as presented in Figs. 8 and 9, where the estimated histograms appear to be different for each one of the phoneme classes. Similar distributions have also been estimated for other phoneme classes. The inter-class differences appear to be quite significant, while only minor ones appear for phonemes of the same class.

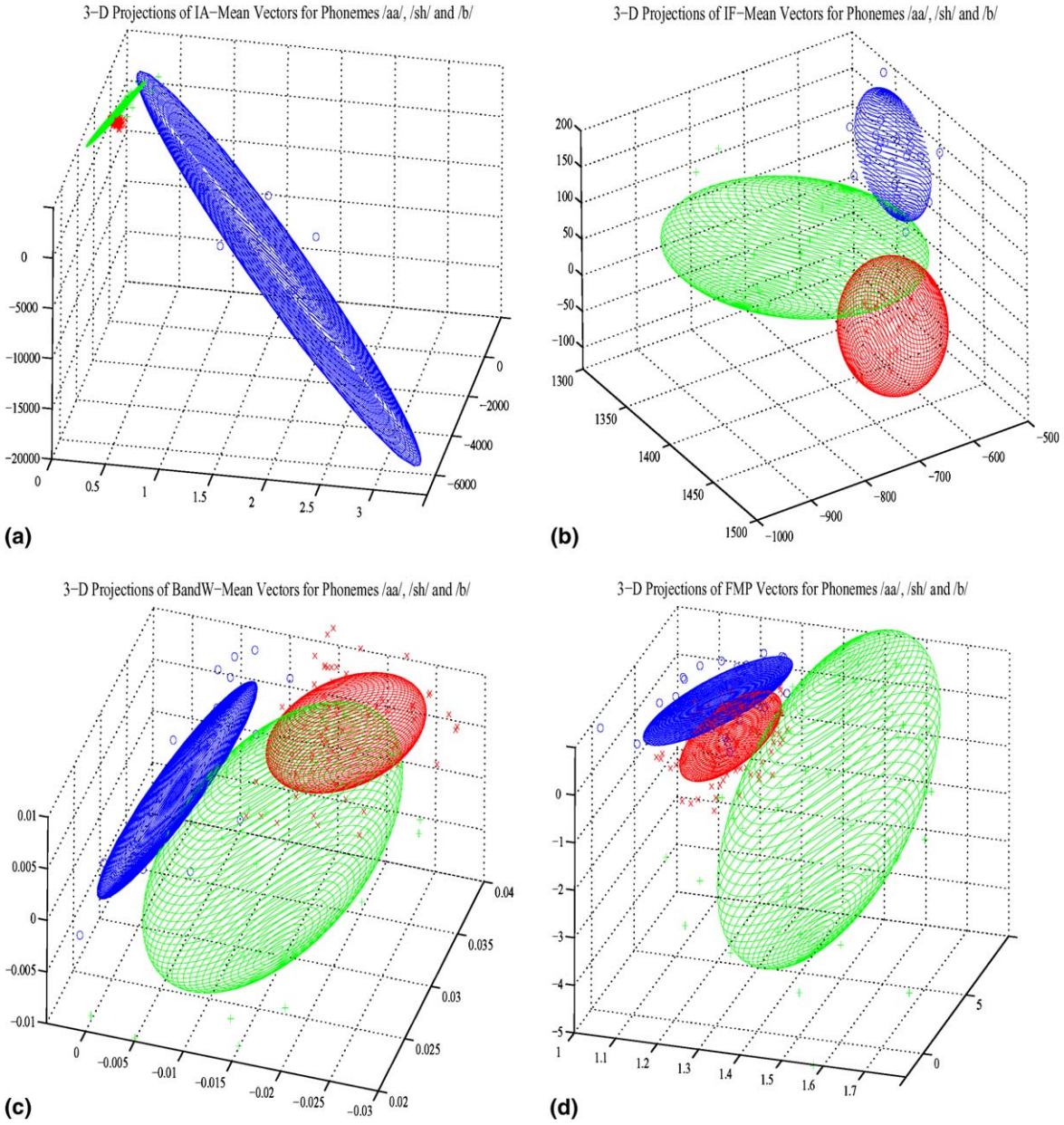

Fig. 14. Instances of 3D DCT projections for phonemes /aa/, /sh/ and /b/of the vectors (a) IA-Mean, (b) IF-Mean, (c) BandW-Mean and (d) FMP. (The markers "o", "x" and "+" represent the projections of the phoneme classes /aa/, /sh/ and /b/, respectively.)

Table 2
Correct phoneme accuracies (%) and mean relative improvement for modulation features on the TIMIT and TIMIT + Noise tasks

Features	Database			
	TIMIT	TIMIT + white	TIMIT + pink	Mean rel. improv.
Phoneme accuracy for the TIMIT tasks (%) (for SNR = 10 dB)				
MFCC	58.40	17.72	18.60	–
MFCC + IA-Mean	59.61	26.03	31.05	23.22
MFCC + IF-Mean	59.34	25.38	30.92	22.11
MFCC + BandW-Mean	59.23	24.78	28.55	18.85
MFCC + FMP	59.92	26.15	32.84	25.56

The distributions of the instantaneous frequency signals appear to follow an exponential law. Their linear parts (when the y-axes are log-scaled) are clear indications of such type of distribution.

In addition, nonlinear features based on the AM–FM resonance model of speech are proposed. Similarly, differences appear in their distributions, as shown in Figs. 10–13. The corresponding mean values and variances appear different and highly dependent on the phoneme classes. Similar conclusions are drawn by studying the 3D ellipsoids which are based on their lower-order statistical moments.

The proposed features are projected to the 3D feature-space and their corresponding equal-likelihood ellipses are estimated, taking under consideration their statistical properties. The corresponding ellipsoid volumes seem to have a relatively small overlap when considering different pairs of phoneme classes. We provide some experimental evidence that the AM–FM features exhibit different and reasonably separable distributions for different phoneme classes.

These experiments show that the proposed features could efficiently be used for speech classification and recognition tasks, as well. They appear to efficiently differentiate for different phoneme classes and thus, they could be useful to adequately discriminate these phoneme classes. Motivated by this feature analysis, we have also applied these AM–FM features to some clean and noisy speech recognition tasks, with a clear improvement of the correct accuracy results when compared to the corresponding results of the MFCCs.

References

Aldroubi, A., Unser, M., Eden, M., 1992. Cardinal spline filters: stability and convergence to the ideal sinc interpolator. Signal Process. 28, 127–138.

Bovik, A.C., Maragos, P., Quatieri, T.F., 1993. AM–FM energy detection and separation in noise using multiband energy operators. IEEE Trans. Signal Process. 41 (Dec.).

Dimitriadis, D., Maragos, P., 2001. An improved energy demodulation algorithm using splines. In: Proc. ICASSP-01, Salt Lake, UT, May.

Dimitriadis, D., Maragos, P., 2003. Robust energy demodulation based on continuous models with application to speech recognition. In: Proc. Eurospeech-03, Geneva, September.

Dimitriadis, D., Maragos, P., Potamianos, A., 2002. Modulation features for speech recognition. In: Proc. ICASSP-02, Orlando, FL, May.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification and Scene Analysis. Wiley, New York.

Fertig, L.B., McClellan, J.H., 1996. Instantaneous frequency estimation using linear prediction with comparisons to the DESAs. IEEE Signal Process. Lett. 3, 54–56.

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.

Kaiser, J.F., 1990. On a simple algorithm to calculate the 'Energy' of a signal. In: Proc. ICASSP-90, Albuquerque, NM, April, pp. 381–384.

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

Maragos, P., Potamianos, A., 1995. Higher order differential energy operators. IEEE Signal Process. Lett. 2 (Aug.), 152–154.

Papoulis, A., 1962. The Fourier Integral and its Applications. McGraw-Hill, New York, NY.

Potamianos, A., Maragos, P., 1994. A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation. Signal Process. 37 (May), 95–120.

Potamianos, A., Maragos, P., 1996. Speech formant frequency and bandwidth tracking using multiband energy demodulation. J. Acoust. Soc. Amer. 99 (6), 3795–3806.

Potamianos, A., Maragos, P., 1999. Speech analysis and synthesis using an AM–FM modulation model. Speech Communication 28, 195–209.

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, NJ.

Ramalingam, C.S., 1996. On the equivalence of DESA-1a and Prony's method when the signal is a sinusoid. IEEE Signal Process. Lett. 3 (May), 141–143.

Unser, M., Aldroubi, A., Eden, M., 1993. B-Spline signal processing: Part I – Theory. Part II – Efficient design and applications. IEEE Trans. Signal Process. 41 (Feb.), 821–848.

Unser, M., Aldroubi, A., Eden, M., 1991. Fast B-Spline transforms for continuous image representation and interpolation. IEEE Trans. Pattern Anal. Machine Intell. 13 (March), 277–285.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002The HTK Book (for HTK Version 3.2). Cambridge Research Lab: Entropics, Cambridge, England.