A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation

Alexandros Potamianos and Petros Maragos

School of Electrical & Computer Engineering, Georgia Institute of Technology,
Atlanta, GA 30332, USA [0]

January 17, 1995

Suggested keywords: *demodulation, energy operator, Hilbert transform, speech processing*

Pages: 41

Figures: 12

Tables: 3

CHANGE OF ADDRESS

Please address all correspondence to:

Prof. Petros Maragos,

School of Electrical & Computer Engineering,

Georgia Institute of Technology,

Atlanta, GA 30332-0250,

U.S.A.

tel: 404-894-3930

fax: 404-894-8363

e-mail: maragos@eedsp.gatech.edu

[0] When this paper was first submitted the authors were at the Division of Applied Sciences, Harvard University, Cambridge, MA 02138, USA. They are currently at the School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

Abstract

[1]

The Hilbert transform together with Gabor's analytic signal provides a standard linear integral approach to estimate the amplitude envelope and instantaneous frequency of signals with a combined amplitude modulation (AM) and frequency modulation (FM) structure. An alternative recent approach uses a nonlinear differential 'energy' operator to track the energy required to generate an AM–FM signal and separate it into amplitude and frequency components. In this paper, we compare these two fundamentally different approaches for demodulation of arbitrary signals and of speech resonances modeled by AM–FM signals. The comparison is done from several viewpoints: magnitude of estimation errors, computational complexity, and adaptability to instantaneous signal changes. We also propose a refinement of the energy operator approach that uses simple binomial convolutions to smooth the energy signals. This smoothed energy operator is compared to the Hilbert transform on tracking modulations in speech vowel signals, band-pass filtered around their formants. The effects of pitch periodicity and band-pass filtering on both demodulation approaches are examined and an application to formant tracking is presented.

The results provide strong evidence that the estimation errors of the smoothed energy operator approach are similar to that of the Hilbert transform approach for speech applications, but smaller for communications applications. In addition, the smoothed energy operator approach has smaller computational complexity and faster adaptation due to its instantaneous nature.

[1]This research work was supported by the National Science Foundation under Grant MIP-91-20624, by a NSF Presidential Young Investigator Award under Grant MIP-86-58150 with matching funds from Xerox, and by a Wallace Fellowship.

1 Introduction

Information in communication systems is usually stored in signals that have a combined amplitude modulation (AM) and frequency modulation (FM) structure. Recently, such signals have been used in [14, 15, 12] to model time-varying amplitude and frequency patterns in speech resonances. Real-valued *AM–FM* signals can be represented as

$$x(t) = a(t)\cos(\underbrace{\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \phi(0)}_{\phi(t)}) \tag{1}$$

Thus $x(t)$ is a cosine of carrier frequency ω_c, with a time-varying amplitude signal $a(t)$ and a time-varying instantaneous angular frequency signal

$$\omega_i(t) \triangleq \frac{d\phi}{dt}(t) = \omega_c + \omega_m q(t), \tag{2}$$

where $q(t) \in [-1, 1]$ is the frequency modulating signal, $\omega_m \in [0, \omega_c]$ is the maximum frequency deviation, and $\phi(0)$ is an arbitrary phase offset.

A typical demodulation problem is, given $x(t)$ and ω_c, to estimate the *amplitude envelope* $|a(t)|$ and *instantaneous frequency* $\omega_i(t)$. A standard approach to this problem is to use the Hilbert transform and the related Gabor's analytic signal [6]; this is well explained in many books on communications or signal processing, e.g., [24, 18]. An alternative approach, recently developed by Maragos, Kaiser, and Quatieri [11, 12], uses an 'energy-tracking' operator to first estimate the energy required for generating the AM–FM signal and then separate it into its amplitude and frequency components. This operator is defined for continuous-time signals $s(t)$ as

$$\Psi_c[s(t)] \triangleq [\dot{s}(t)]^2 - s(t)\ddot{s}(t) \tag{3}$$

where $\dot{s} = ds/dt$. Its counterpart for discrete-time signals $s(n)$ is

$$\Psi_d[s(n)] \triangleq s^2(n) - s(n-1)s(n+1) \tag{4}$$

The nonlinear operators Ψ_c and Ψ_d were developed by Teager during his work on speech production modeling [25, 26] and were first introduced systematically by Kaiser [8, 9]. When Ψ_c is applied to signals produced by a simple harmonic oscillator, e.g. a mass-spring oscillator, it can track the oscillator's energy (per half unit mass), which is equal to the squared product of the oscillation amplitude and frequency. Thus we henceforth refer to Ψ_c, Ψ_d as the *energy operators*. The energy operator approach to demodulation has many attractive features such as simplicity, efficiency, and adaptability to instantaneous signal variations [11, 12].

In this paper we compare these two fundamentally different approaches to AM–FM signal demodulation. The Hilbert transform approach mainly involves a linear integral operator, whereas the energy operator approach uses a nonlinear differential operator. Both are briefly reviewed in Sections 2.1 and 2.2. In Section 2.3 we introduce an improvement of the energy operator approach where the energy signals undergo some smoothing before being used for demodulation. We refer to the above three approaches to AM–FM demodulation as amplitude/frequency *separation algorithms.* In Section 3 we provide detailed comparisons among these separation algorithms, applied to arbitrary synthetic signals, from many different viewpoints: magnitude of estimation errors, computational complexity, behavior in the presence of noise, and adaptability rate in the presence of abrupt signal changes. Most of these issues are discussed on an experimental basis.

A promising application area for the methods compared in this paper is the problem of tracking modulations in speech resonances, keeping in mind the importance of formants in speech processing [23]. Motivated by several nonlinear and time-varying phenomena during speech production, Maragos, Quatieri, and Kaiser [14, 15] proposed an AM–FM modulation model for the production of speech signals, by representing a single speech resonance (formant) within a pitch period as a damped AM–FM signal

$$R(t) = a(t)\cos(\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \theta) \tag{5}$$

where ω_c is the center value of the formant frequency, $q(t)$ is the normalized frequency modulating signal, ω_m is the maximum frequency deviation from ω_c, and $a(t) = e^{-\sigma t}A(t)$ is a time-varying amplitude that includes an exponential decay and a generally non-constant signal $A(t)$. The instantaneous value of the formant frequency is $\omega_i(t) = \omega_c + \omega_m q(t)$. Finally, the speech signal $S(t)$ within a pitch period is modeled as the sum $S(t) = \sum_{k=1}^{N} R_k(t)$ of N such AM–FM signals, where N is the number of speech formants.

The AM–FM modulation model and the energy separation algorithm have been used successfully for determining the center values of the formant frequencies in a speech segment [7]. Another application is the AM–FM modulation vocoder [7], that extracts the *formant bands* from the spectrum using a bank of adaptive Gabor filters. The formant bands are then demodulated to amplitude envelope and instantaneous frequency, decimated and coded. At the receiver, the speech bands are reconstructed from the modulating signals and added together. Finally, in [5] the energy operator is used to determine formant tracks used for stop phoneme classification.

Before applying either demodulation approach to a single speech resonance signal, one needs first to isolate the resonance through *band-pass filtering* of the speech signal. Within certain constraints on the impulse response of the band-pass filters, Papoulis [19, ch.8] (in his discussion

on spectrum analyzers) and Flanagan [4] (in his work on coding of speech spectra) have shown that the effect of band-pass filtering can be expressed as a combined amplitude and phase modulation, whose components are identical to the the short-time Fourier transform magnitude and phase extracted via the Hilbert transform [4]. Another effect of band-pass filtering is to blur the true input modulating signals. In Section 4.2, following the work of Papoulis on this subject [19, ch.7], we analyze the blurring imposed by the filter on a general AM–FM input signal. In addition, speech vowel signals have the *pitch periodicity* which poses certain problems to AM–FM modeling and demodulation. Hence, in Section 4 we compare experimentally the effects of these two unavoidable problems (i.e., band-pass filtering and pitch periodicity) when applying the Hilbert transform and energy operator demodulation approaches to speech vowels. Also, to demonstrate the underlying ideas from our experiments on real speech more clearly, several comparisons on *synthesized* speech vowels are presented in Section 4. We have found that the conclusions drawn from experiments on synthetic and real speech are similar. Further, since the parameters of the synthetic speech signals are known, the accuracy of the various demodulation approaches can be checked more easily. We conclude Section 4 with an application of the energy separation algorithm and the AM–FM modulation model to formant tracking. Three new algorithms are proposed, two that use an iterative demodulation approach and one with a multiband parallel architecture. Finally, in Section 5 we summarize our conclusions from the comparisons of the two AM–FM demodulation approaches.

2 Amplitude/Frequency Separation Algorithms

Consider a real-valued AM–FM signal $x(t) = a(t)\cos[\phi(t)]$ as in (1). By 'amplitude/frequency separation' or 'demodulation' we shall henceforth refer to the estimation of the amplitude envelope $|a(t)|$ and instantaneous frequency $\omega_i(t) = \dot{\phi}(t)$, which we call *information signals*. Assuming that ω_c is known, estimating $\omega_i(t)$ is equivalent to estimating the frequency modulating signal $q(t)$. Similarly, if $a(t) \geq 0$ for all t, one can write $a(t) = A[1+\kappa b(t)]$ where $0 \leq \kappa \leq 1$, $|b(t)| \leq 1$, and A is some constant; estimating $|a(t)|$ is equivalent to estimating the amplitude modulating signal $b(t)$. In general, amplitude signals $a(t)$ may be nonnegative, however, for the purposes of this paper we consider it sufficient to estimate the amplitude envelope.

2.1 Hilbert Transform Separation Algorithm

The *Hilbert transform* of any AM–FM signal $x(t) = a(t)\cos[\phi(t)]$ is

$$\hat{x}(t) = x(t) * \frac{1}{\pi t} \tag{6}$$

with Fourier transform

$$\hat{X}(\omega) = -j\ \mathrm{sgn}(\omega)\, X(\omega) \tag{7}$$

where $X(\omega)$ is the Fourier transform of $x(t)$. Given the *analytic signal*

$$z(t) = x(t) + j\hat{x}(t) = r(t)\ \exp[j\theta(t)] \tag{8}$$

its modulus $r(t)$ and phase derivative $\dot{\theta}(t)$ can serve as (generally approximate) estimates for the amplitude envelope and instantaneous frequency of $x(t)$. Thus the *Hilbert transform separation algorithm (HTSA)* is given by the following two equations:

$$r(t) \;=\; \sqrt{x^2(t) + \hat{x}^2(t)} \;\approx\; |a(t)| \tag{9}$$

$$\dot{\theta}(t) \;=\; \frac{d}{dt}(\arctan\left[\frac{\hat{x}(t)}{x(t)}\right]) \;\approx\; \omega_i(t) \tag{10}$$

Let the *quadrature signal* of $x(t)$ be defined as

$$x_q(t) = a(t)\sin[\phi(t)] \tag{11}$$

Clearly, if the Hilbert transform of $x(t)$ is equal to its quadrature signal, then the HTSA estimates $r(t)$ and $\dot{\theta}(t)$ are equal to the actual information signals $|a(t)|$ and $\omega_i(t)$. In general though, $\hat{x}(t)$ and $x_q(t)$ are not equal, thus an envelope $e_a(t)$ and frequency $e_\omega(t)$ estimation error is present. These estimation errors are closely related to the *quadrature error* signal of the Hilbert transform defined as

$$e(t) = \hat{x}(t) - x_q(t) = \hat{x}(t) - a(t)\sin[\phi(t)] \tag{12}$$

Consider the complex-valued signal

$$w(t) = x(t) + jx_q(t) = a(t)\exp[j\phi(t)] \tag{13}$$

with Fourier transform $W(\omega)$. Then

$$X(\omega) = \frac{1}{2}[W(\omega) + W^*(-\omega)] \tag{14}$$

Nuttal [16, 17] has shown that the total energy in the quadrature error signal is

$$E = \int_{-\infty}^{+\infty} |e(t)|^2 dt = \frac{1}{\pi}\int_{-\infty}^{0} |W(\omega)|^2 d\omega \tag{15}$$

Therefore, if $W(\omega)$ is zero for negative frequencies the quadrature error $e(t)$ will also be zero. When $W(\omega)$ extends to negative frequencies, the quadrature error $e(t)$ is non-zero and the magnitude of the error increases as the negative side of $W(\omega)$ grows. In the special case of a *cosine* $x(t) = A\cos(\omega_c t)$, the HTSA provides exact estimates of the amplitude and frequency, because $\hat{x}(t) = A\sin(\omega_c t)$ and hence $e(t) = 0$ for all t.

The quadrature error $e(t)$ can provide bounds for the estimation errors of the information signals. The envelope estimation error $e_a(t)$ may be expressed as:

$$\begin{aligned} e_a(t) &= |a(t)| - r(t) = |a(t)| - \sqrt{x^2(t) + \hat{x}^2(t)} \\ &= |a(t)| - \sqrt{x^2(t) + [x_q(t) + e(t)]^2} \\ &= |a(t)| \left(1 - \sqrt{1 + 2\frac{e(t)}{a(t)}\sin[\phi(t)] + \frac{e^2(t)}{a^2(t)}}\right) \\ &\approx -e(t)\,\mathrm{sgn}[a(t)]\sin[\phi(t)] - \frac{e^2(t)}{2|a(t)|}, \quad \text{if} \quad |e(t)| \ll |a(t)|,\ a(t) \neq 0 \end{aligned} \tag{16}$$

The phase estimation error is

$$\begin{aligned} e_\phi(t) &= \arctan\left[\frac{\hat{x}(t)}{x(t)}\right] - \phi(t) \\ &= \arctan\left[\,\tan[\phi(t)] + \frac{e(t)}{a(t)\cos[\phi(t)]}\,\right] - \phi(t) \end{aligned}$$

Hence:

$$\begin{aligned} \tan[e_\phi(t) + \phi(t)] = \tan[\phi(t)] + \frac{e(t)}{a(t)\cos[\phi(t)]} \quad &\Longrightarrow \\ \tan[e_\phi(t)] = \frac{e(t)\cos[\phi(t)]}{a(t) + e(t)\sin[\phi(t)]} \quad &\Longrightarrow \\ e_\phi(t) \approx \frac{e(t)\cos[\phi(t)]}{a(t)}\,, \qquad \text{if} \quad |e(t)| \ll |a(t)|, \ \ a(t) \neq 0 \end{aligned} \tag{17}$$

Finally, the instantaneous frequency estimation error e_ω is simply the derivative of the phase error e_ϕ. In brief, for $|e(t)| \ll |a(t)|$ and $a(t) \neq 0$

$$|e_a(t)| \leq |e(t)| \tag{18}$$

$$|e_\phi(t)| \leq \left|\frac{e(t)}{a(t)}\right| \tag{19}$$

$$|e_\omega(t)| = |\dot{e}_\phi(t)| \approx \left|\frac{d}{dt}\left[\frac{e(t)\cos[\phi(t)]}{a(t)}\right]\right| \tag{20}$$

For discrete-time signals $x(n)$ their Hilbert transform $\hat{x}(n) = x(n) * h(n)$ is defined [18] in the time domain as the convolution of $x(n)$ with the infinite impulse response

$$h(n) = \begin{cases} \frac{2}{\pi}\frac{\sin^2(\pi n/2)}{n} & n \neq 0 \\ 0 & n = 0 \end{cases} \tag{21}$$

In practice, one can implement the Hilbert transform by using an FIR approximation to the IIR $h(n)$. Such FIR filter designs can be obtained either via the window method (e.g. Kaiser windows) or the equiripple method [18]. An alternative way of approximating the discrete-time analytic signal $z(n) = x(n) + j\hat{x}(n)$ is by using FFTs to implement a 90^o phase splitter [18].

If $x(n) = a(n)\cos[\phi(n)]$ and $x_q(n) = a(n)\sin[\phi(n)]$, the total energy of the quadrature error becomes in discrete time

$$E = \sum_{n=-\infty}^{\infty} |\hat{x}(n) - x_q(n)|^2 = \frac{1}{\pi}\int_{-\pi}^{0} |W(\Omega)|^2 d\Omega \tag{22}$$

where $W(\Omega)$ is the discrete Fourier transform of the signal $w(n) = x(n) + jx_q(n)$. The envelope and instantaneous frequency estimation equations (9),(10) and error bound equations (18),(19),(20) also hold in discrete time. However, in addition to the quadrature error that depends on the signal $x(n)$, any (FIR or FFT) discrete Hilbert transform implementation also incurs an additional error, by being an approximation of the exact IIR Hilbert transformer.

In this paper, we will use two FIR Hilbert transformers designed via the window method: (i) a filter with a short 19-sample impulse response, cutoff frequency at 500 Hz, and 10% maximum ripple in the passband; (ii) a filter with a long 139-sample impulse response, cutoff frequency at 200 Hz, and 1% ripple. The two Hilbert transform separation algorithms corresponding to the above implementations will be referred to as 'short HTSA' and 'long HTSA', respectively. Both implementations assume a sampling frequency of 20 KHz.

2.2 Energy Operator Separation Algorithm

In [15, 13] it has been shown that, when Ψ_c is applied to an AM–FM signal

$$x(t) = a(t)\cos[\int_0^t \omega_i(\tau)d\tau]$$

it can approximately estimate the squared product of the amplitude and frequency signals; i.e.,

$$\Psi_c[x(t)] \approx [a(t)\omega_i(t)]^2 \tag{23}$$

assuming that the signals $a(t)$ and $\omega_i(t)$ do not vary too fast (rate of change) or too greatly (range of value) with time compared to the carrier frequency ω_c. For the demodulation of AM–FM signals $x(t)$, the following *energy operator separation algorithm (EOSA)* has been developed in [11, 12]:

$$\sqrt{\frac{\Psi_c[\dot{x}(t)]}{\Psi_c[x(t)]}} \approx \omega_i(t) \tag{24}$$

$$\frac{\Psi_c[x(t)]}{\sqrt{\Psi_c[\dot{x}(t)]}} \approx |a(t)| \tag{25}$$

At each time instant the EOSA estimates the instantaneous frequency and the amplitude envelope of x, by using the output values of the energy operator applied to the signal x and the signal derivative $\dot{x}$. Upper bounds for the approximation errors in (23) and (24),(25) have been found in [15, 12, 13]; the bounds are expressed in terms of the ratios of the bandwidths of a and ω_i vs. the carrier ω_c. If $x(t) = A\cos(\omega_c t)$ is a cosine with no AM or FM, the EOSA yields exact estimates of the constant amplitude and frequency.

Similar methods can be applied to the discrete-time AM–FM signal

$$x(n) = a(n)\cos[\phi(n)] = a(n)\cos(\Omega_c n + \Omega_m \int_0^n q(k)dk + \phi(0))$$

to estimate its amplitude envelope $|a(n)|$ and instantaneous frequency

$$\Omega_i(n) = \frac{d\phi}{dn}(n) = \Omega_c + \Omega_m q(n)$$

where $0 \leq \Omega_m \leq \Omega_c$ and $|q(n)| \leq 1$. Note that the continuous-time frequencies ω_c, ω_m, and ω_i have been replaced by their discrete-time counterparts Ω_c, Ω_m, and Ω_i, which are assumed to be in $[0, \pi]$. If $x(n)$ has resulted from sampling of a continuous-time signal, then

$$\Omega_c = \omega_c T \quad , \quad \Omega_m = \omega_m T \quad , \quad \Omega_i = \omega_i T$$

where T is the sampling period. It has been shown in [15, 13] that

$$\Psi_d[x(n)] \approx a^2(n)\sin^2[\Omega_i(n)] \tag{26}$$

By applying Ψ_d to both x and its backward difference

$$y(n) = x(n) - x(n-1)$$

a discrete-time EOSA has been developed in [11, 12]:

$$\arccos\left(1 - \frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4\Psi_d[x(n)]}\right) \approx \Omega_i(n) \tag{27}$$

$$\sqrt{\frac{\Psi_d[x(n)]}{1 - \left(1 - \frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4\Psi_d[x(n)]}\right)^2}} \approx |a(n)| \tag{28}$$

$$\tag{29}$$

The frequency estimation part assumes that $0 < \Omega_i(n) < \pi$. Thus, the discrete EOSA algorithm can estimate instantaneous frequencies up to 1/2 the sampling frequency. The approximations in (26) and (27),(28) are valid under assumptions similar to the continuous-time case.

One of the underlying ideas in the EOSA is that of signal estimation by using nonlinear combinations of the 'instantaneous' values of the signal and its derivatives. From this viewpoint, it is related to the signal modeling framework of [28]. However, the usage of an energy operator puts forth the additional interesting intuitive feature of tracking the energy required for generating the AM–FM signal and then separating it into amplitude and frequency components.

2.3 Smoothed Energy Operator Separation Algorithm

The precise result from applying the energy operator to an AM–FM signal is [11]:

$$\Psi_c[a(t)\cos(\phi(t))] = \overbrace{(a\dot{\phi})^2}^{D} + \overbrace{\frac{\Psi_c(a)}{2}}^{E_L} + \overbrace{a^2\ddot{\phi}\sin(2\phi) + \frac{\Psi_c(a)}{2}\cos(2\phi)}^{E_H} \tag{30}$$

The desired term is $D = (a\dot{\phi})^2$, whereas E_L and E_H are the error terms in the approximation (23). Note that the energy operator approximation incurs a low-frequency error component E_L and a high-frequency component E_H concentrated around $2\,\omega_c$, that is twice the carrier frequency of the AM–FM signal. In addition, the desired term D is bandlimited with a highest frequency that is much smaller than ω_c (under the assumption that the bandwidth of the amplitude $a(t)$ and the instantaneous frequency $\dot{\phi}(t)$ signals is also much smaller than ω_c). This means that the high-frequency error component is well separated from the desired term in the frequency domain. Thus, by filtering the energy operator output through an appropriate low-pass filter, one can eliminate the high-frequency error component without affecting the low-frequency desired term. Fig. 1

Similarly, in discrete-time, when the energy operator Ψ_d is applied to an AM–FM signal, we get a high-frequency error component concentrated around $2\,\Omega_c$ as shown in Figs. 1 (a),(b). The total error signal of the approximation and its spectrum are displayed for an AM–FM/Cosine signal (50% AM, 20% FM) with carrier frequency at $\Omega_c = 0.2\pi$. Clearly, the error has a high frequency component around $2\,\Omega_c = 0.4\pi$ that can be eliminated through low-pass filtering.

The choice of an appropriate low-pass filter is not straightforward. Clearly, an 'expensive' filter with a long impulse response can decrease considerably the approximation error. We must keep in mind though, that one of the major advantages of the energy operator is its instantaneous nature, which guarantees excellent time resolution. This property is valuable for applications where the information signals may have abrupt transitions (e.g., pitch or phoneme transitions in speech analysis). To conserve the instantaneous nature and the simplicity of Ψ_d one can choose an FIR filter with a short impulse response. For our purposes we will use a 7–point linear binomial smoothing filter with impulse response (1, 6, 15, 20, 15, 6, 1). [2] This filter is equivalent to the 3-point filter

[2] Smoothing the output of the energy operator via low-pass filtering to reduce estimation errors has also been done

(1, 2, 1) applied to the energy operator output three times or to the 2-point moving average filter (1, 1) applied six times. With this simple and computationally inexpensive smoothing, the energy operator approximation error decreases typically by 50%. Also, the envelope and frequency estimation errors are reduced when the smoothed energy signals are used in the separation algorithm. Henceforth, we will refer to the envelope and frequency separation algorithm using the binomially smoothed energy signals, as the *Smoothed Energy Operator Separation Algorithm (SEOSA)*, summarized in Fig. 2. Fig. 2

It is shown in [13] that the discrete-time operator Ψ_d results from the continuous-time one Ψ_c by approximating $\dot{x}(t)$ by $x(n)-x(n-1)$. In addition, Ψ_d followed by a 3-point binomial filter (1, 2, 1) is equivalent to using the 3-sample symmetric difference $[x(n+1)-x(n-1)]/2$ for approximating $\dot{x}(t)$. Approximations of $\dot{x}(t)$ that involve more samples (longer derivatives) offer an alternative way of improving the performance of the energy operator, with results similar to the binomial filter smoothing.

Finally, smoothing can be applied on the estimated information signals (post-smoothing) instead of the energy signals (pre-smoothing), as the envelope and frequency estimation error signals have a high frequency component around $2\omega_c$, very much like the energy operator error does. Both approaches (post and pre-smoothing) yield similar results. There are applications though where post-smoothing is advantageous (e.g. when median filtering is also performed).

3 Comparisons on Synthetic Signals

The three amplitude/frequency separation algorithms (EOSA, SEOSA, and HTSA) are compared here using discrete-time AM–FM/Cosine signals $x(n)$ of the form:

$$x(n) = (1 + \kappa \cos(\Omega_a n)) \cos[\Omega_c n + (\Omega_m/\Omega_f) \sin(\Omega_f n)] \tag{31}$$

The corresponding sinusoidal amplitude and instantaneous frequency signals are:

$$a(n) = 1 + \kappa \cos(\Omega_a n) \tag{32}$$

$$\Omega_i(n) = \Omega_c + \Omega_m \cos(\Omega_f n) \tag{33}$$

The AM modulation index $\kappa \in (0,1)$ and the FM modulation depth $\Omega_m/\Omega_c \in (0,1)$ determine, respectively, the amount of AM and FM. An example of estimating the amplitude envelope and the instantaneous frequency using the long HTSA (139-sample FIR design), the EOSA and SEOSA is

in [21] when using the energy operator as detector of transient signal signatures in AM–FM background noise.

shown in Figs 3 and 4 for an AM–FM/Cosine signal with modulation amounts of 60% AM ($\kappa = 0.6$) and 30% FM ($\Omega_m/\Omega_c = 0.3$). We observe that all three algorithms yield good estimates for the amplitude envelope and instantaneous frequency signals. A few small ripples found in the EOSA estimates are eliminated via the simple binomial smoothing involved in the SEOSA. Fig. 3, 4

From extensive experiments on the class of AM–FM/Cosine signals the performance of the separation algorithms was found to depend mainly upon the ratio $\Omega_c/\Omega_{a,f}$ where $\Omega_{a,f} = \max(\Omega_a, \Omega_f)$ (i.e., the ratio of the carrier frequency over the bandwidth of the information signals), the AM index κ, and the FM depth Ω_m/Ω_c. Henceforth in this section we assume that $\Omega_{a,f} = \Omega_a = \Omega_f$. Extensive numerical comparisons among the HTSA [using both the short FIR (19 samples) and the long (139 samples) implementations], the EOSA and SEOSA on AM–FM/Cosine signals were performed by varying these three parameters. A typical value for the ratio $\Omega_c/\Omega_{a,f}$ for speech analysis applications is about 10, since an average formant value is at 2 KHz and the bandwidths of the amplitude/frequency modulating signals have been found at 200 Hz on the average. However, in communication systems the ratio $\Omega_c/\Omega_{a,f}$ takes much higher values; e.g., in AM radio the ratio is in the order of a 100, whereas in FM radio it is in the order of 1000. For our experiments, the case where $\Omega_c/\Omega_{a,f} = 10$ will be referred to as *'speech specifications'* and the case where $\Omega_c/\Omega_{a,f} \geq 100$ will be referred to as *'communications specifications'*. Table 1-3

Table 1 shows the mean absolute error for envelope and frequency estimation *averaged* over 100 different AM–FM/Cosine signals with AM index varying from 5–50% (step 5%) and FM depth varying from 2-20% (step 2%). The carrier frequency was fixed at $\Omega_c = \pi/5$. The average errors are displayed for all four separation algorithms for $\Omega_c/\Omega_{a,f} = 10$ and 100. Overall the long HTSA yielded approximately one order of magnitude smaller error than the EOSA and SEOSA for speech specifications. Yet, the short HTSA (with approximately the same computational complexity as the EOSA) performed much worse than the EOSA. The smoothed EOSA estimation error is 30-50% smaller than the error of the EOSA without smoothing. For communications specifications, the errors of both the EOSA and SEOSA decrease by one order of magnitude, becoming comparable or somewhat smaller to that of the long HTSA. Note that for $\Omega_c/\Omega_{a,f} = 1000$ the EOSA error becomes one order of magnitude smaller than the long HTSA error.

An important issue is how does the performance of the separation algorithms deteriorate in the presence of noise. Table 2 shows the mean absolute estimation error (envelope and frequency) averaged over the same 100 cases of AM–FM/Cosine signals used in Table 1, in the presence of added white Gaussian noise at a signal-to-noise ratio (SNR) of 30 db. For both speech and communications specifications, the long HTSA performs better than the EOSA. This is expected,

since the HTSA involves an integral transform that does implicit smoothing, whereas the EOSA involves a an 'instantaneous' differential operator. Interestingly, the smoothed EOSA (which uses the simple 7-point binomial smoother) yields an error comparable or smaller to the long HTSA error. Finally, we note that all our comparisons of the estimation errors (both in the noise-free and noise-corrupted case) are based on numerical simulations and refer to the special case of AM–FM/Cosine signals. A theoretical analysis of the HTSA approach for random signals can be found in [20], and a theoretical analysis for the performance of the EOSA in the presence of Gaussian noise has been recently developed in [2].

Table 3 shows that out of the four algorithms the EOSA has the smallest computational complexity and needs the smallest number W of input samples per single output estimate in its moving window. The smoothed EOSA needs a few more additions (due to the binomial smoothing) and a window twice as long. The short HTSA has similar computational complexity with the EOSA but needs a four times longer window. Finally, the long HTSA has about one order of magnitude higher computational complexity than the EOSA and 3-4 times higher than the SEOSA. The biggest drawback of the long HTSA is that it requires a window more than one order of magnitude wider than the window of the EOSA and SEOSA. Hence, the EOSA and SEOSA have the advantage of adapting instantaneously and needing an extremely small number of input samples to operate. Analysis of speech signals using one of the amplitude/frequency separation algorithms is done on a short-time basis. From speech analysis experiments, we observed that the long HTSA implementation needs an FIR length $W \approx N$, where N is the average length of the short-time speech analysis frame. Hence, in this case, the computational complexity of the HTSA is quadratic $O(N^2)$. In contrast, the complexity of the EOSA and SEOSA is always linear $O(N)$, and the multiplicative constant is very small.

Fig. 5

Finally, in Fig. 5 we compare the amplitude envelope and instantaneous frequency estimation errors for the three algorithms (EOSA, SEOSA and long HTSA), for values of the ratio $\Omega_c/\Omega_{a,f}$ ranging from 10 to 1000 (Ω_c is set to $\pi/5$ throughout the experiment). Each point in the error curves represents the average error over 100 experiments for AM–FM/Cosine signals with a varying AM index $\kappa \in [0.05, 0.5]$ and a varying FM depth $\Omega_m/\Omega_c \in [0.02, 0.2]$ (as in Table 1). We observe that the EOSA error is decreasing linearly with the ratio $\Omega_c/\Omega_{a,f}$, while the HTSA error remains approximately constant. For $\Omega_c/\Omega_{a,f} = 10$ the EOSA error is one order of magnitude larger than the HTSA error. As the ratio approaches 100 the EOSA and HTSA error are of comparable magnitude. Finally, when the ratio is in the neighborhood of 1000 the EOSA error is one order of magnitude smaller. The SEOSA error decreases linearly with $\Omega_c/\Omega_{a,f}$ and is approximately

half of the EOSA error. Note though, that the SEOSA envelope error reaches a satiation point for ratio values over 400 and then it becomes larger than the EOSA error. This happens because the binomial smoothing in the SEOSA eliminates part of the EOSA approximation error, but it also degrades slightly the desired energy terms since it uses a non-ideal low-pass filter. Thus, for very large values of the ratio $\Omega_c/\Omega_{a,f}$ the EOSA approximation error becomes smaller than the information signal degradation introduced by the SEOSA.

4 Experiments on Real Speech

In this section, the above presented separation algorithms will be applied to real and synthetic speech signals to track the envelope and the frequency modulation of speech resonances. A single speech resonance is modeled as an exponentially damped AM–FM signal (5); the speech signal is the sum of such AM–FM signals.

Before applying the demodulation algorithms to a speech resonance, we must first extract the resonance through band-pass filtering. For this purpose, we will use a Gabor filter (for reasons presented in [11]), with impulse and frequency response:

$$h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t) \tag{34}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}\left(\exp\left[-\frac{(\omega-\omega_c)^2}{4\alpha^2}\right] + \exp\left[-\frac{(\omega+\omega_c)^2}{4\alpha^2}\right]\right) \tag{35}$$

The center frequency ω_c of the filter is chosen equal to the center formant frequency. The parameter α controls the bandwidth of the Gabor filter; the effective bandwidth of the Gabor filter is equal to $\alpha/\sqrt{2\pi}$. Note that in discrete time the impulse response $h(n)$ is a sampled version of (34).

In addition to the need of band-pass filtering (to extract a single resonance), voiced speech signals also have the feature of the pitch periodicity. Henceforth, we will examine the performance of the long HTSA and the SEOSA for speech resonance demodulation. In particular, we will investigate how the separation algorithms are affected by the pitch periodicity, by band-pass filtering and by speech transitions.

4.1 Effects of Pitch

Fig. 6

The effects of pitch on envelope and frequency estimation are studied first on a synthesized signal $s(n)$, modeling a single speech formant. In our example, $s(n)$ is the output of a linear time-invariant speech resonator with a single resonance at 1300 Hz (bandwidth = 30 Hz), excited by a periodic sequence of unit pulses with (pitch) frequency at 100 Hz. Fig. 6 (a),(b) show the signal

$s(n)$ and the excitation. The amplitude envelope and instantaneous frequency estimates for the energy operator and the Hilbert transform separation algorithms are shown at Fig. 6 (c)-(f). The HTSA envelope estimate (c) shows clearly the exponential decay of the actual envelope, yet it also displays misleading modulations around the instants that the pitch pulses occur. Similarly, the HTSA instantaneous frequency estimate (d) tracks correctly the formant frequency at 1300 Hz, but in the neighborhood of the pitch pulses the frequency estimate is heavily modulated. The SEOSA estimated envelope (e) consists of decaying exponentials interrupted by a small spike at each pitch pulse. In the same way, the instantaneous frequency SEOSA estimate (f) is constant everywhere at 1300 Hz, except at the location of the pitch pulses where large (double) spikes occur.

In brief, the SEOSA envelope and instantaneous frequency estimates deterioration (due to the pitch) is concentrated at the instants when the pitch pulses occur, while for the HTSA the estimation error is significant in a time interval of 3-5 msec around that instants. Clearly, the HTSA by using an integral transform does implicit smoothing (low-pass filtering) to the information signals. Thus, the high frequency component of the event (pitch pulse) is filtered out and what we see in our plots is a low-pass filtered spike (with modulations). On the other hand, the energy operator is an 'instantaneous' differential operator, whose discrete implementation involves a very short analysis window (a very few input samples per output sample). As a result, the SEOSA estimates have superior time resolution than the HTSA ones (e.g. abrupt transitions are better preserved).

Of interest is also the general case of a discontinuity at the envelope or the instantaneous frequency of an AM–FM speech-like signal. Consider for example the case where a jump occurs to the carrier frequency ω_c of the AM–FM signal. As in the previous example, the HTSA estimation breaks down in the neighborhood of the instant when the carrier frequency discontinuity occurs, presenting erroneous modulations (especially in instantaneous frequency estimation). The SEOSA estimates, however, have a (single or double) spike at the jump instant and are 'correct' elsewhere. So, the actual envelope and instantaneous frequency around the discontinuity are easily recoverable from the SEOSA followed by median filtering. In addition, the spiky nature of the energy signals at the instant of the jump will inform us of the transition/event. Thus, the energy operator can serve as an event detector (as opposed to the Hilbert transform which tends to smooth out discontinuities). An application of the energy operator event detector property in underwater acoustics is presented in [22].

Fig. 7

An example of how a discontinuous carrier frequency ω_c affects the instantaneous frequency estimation is shown in Fig 7. The carrier frequency of an AM–FM signal $x(n)$ jumps by 1.43% at the 250th sample, causing a discontinuity to the signal itself. The HTSA and SEOSA frequency

estimates are shown (c, d). Clearly, the HTSA frequency estimate presents erroneous modulations in the neighborhood of the jump, while the SEOSA error is concentrated only in 10-12 samples, around the point of discontinuity.

4.2 Effects of Bandpass filtering

For any AM–FM signal $x(t) = a(t)\cos[\phi(t)]$ we may write:

$$x(t) = a(t)\cos[\omega_c t + p(t)] = a_1(t)\cos[\omega_c t] - a_2(t)\sin[\omega_c t] \tag{36}$$

where

$$a(t) = \sqrt{a_1^2(t) + a_2^2(t)} \qquad p(t) = \arctan\left[\frac{a_2(t)}{a_1(t)}\right] \tag{37}$$

Next, we filter $x(t)$ through a band-pass filter with impulse response:

$$h(t) = h_\ell(t)\cos[\omega_c t] \tag{38}$$

where $h_\ell(t)$ is the impulse response of the corresponding low-pass filter and ω_c is the carrier frequency of the AM–FM signal $x(t)$. Then the filtered signal $\tilde{x}(t) = x(t) * h(t)$ will be given approximately by (the conditions under which the approximation holds can be found in [19, ch. 7]):

$$\tilde{x}(t) = \tilde{a}(t)\cos[\omega_c t + \tilde{p}(t)] \approx \frac{1}{2}\left[a_1(t) * h_\ell(t)\right]\cos[\omega_c t] - \frac{1}{2}\left[a_2(t) * h_\ell(t)\right]\sin[\omega_c t] \tag{39}$$

and the new amplitude envelope and instantaneous frequency will be:

$$\tilde{a}(t) \approx \frac{1}{2}\sqrt{\left[(a(t)\cos[p(t)]) * h_\ell(t)\right]^2 + \left[(a(t)\sin[p(t)]) * h_\ell(t)\right]^2} \tag{40}$$

$$\tilde{\omega}_i(t) = \omega_c + \frac{d}{dt}[\tilde{p}(t)] \approx \omega_c + \frac{d}{dt}\arctan\left[\frac{(a(t)\sin[p(t)]) * h_\ell(t)}{(a(t)\cos[p(t)]) * h_\ell(t)}\right] \tag{41}$$

We conclude that the amplitude envelope and instantaneous frequency of a band-pass filtered AM–FM signal $\tilde{x}(t)$ are low-pass filtered versions of the actual information signals.

4.2.1 Effects of Gabor filtering

Fig. 8

An example of how band-pass filtering affects the SEOSA and HTSA estimation is shown in Fig. 8. The vowel /a/ with formants at 600, 1200, 2500 and 3600 Hz (and bandwidth, respectively, at 30, 50, 80 and 110 Hz) is synthesized using time-invariant linear resonators in cascade, excited by a sequence of unit pulses (pitch frequency at 100 Hz). Then, the synthetic signal is band-pass filtered around its third formant, using a Gabor filter with center frequency at $f_c = 2500$ Hz and bandwidth parameter $\alpha = 1000$ Hz. In Fig. 8 (a), (b) three pitch periods of the synthetic vowel /a/ and the

extracted resonance are shown. The HTSA and SEOSA estimates for the amplitude envelope and the instantaneous frequency of the resonance at 2500 Hz are shown in (c)-(f). Interestingly, both separation algorithms (HTSA, SEOSA) produce almost identical estimates. The amplitude envelope estimates (c), (e) are exponentially decaying with smooth transitions at the instants when pitch pulses occur. The frequency estimates (d), (f) are everywhere equal to the center formant frequency ($f_c = 2500$), apart from 3-5 msec around the pitch pulses, where they deviate considerably from f_c.

From Fig. 6 we know what the envelope and frequency estimates would be if filtering was not necessary to extract the resonance (single formant case). We observe that by band-pass filtering the original signal, double spikes turn into smooth 'sinusoidal' curves (SEOSA frequency estimation (f)) and jumps into smooth transitions (SEOSA envelope estimation (e)). This is anticipated, because band-pass filtering actually filters out the higher frequency components of the envelope and instantaneous frequency signals (see eq. (40), (41)). In that sense, after band-pass filtering of the original signal, the SEOSA and HTSA display similar results, as now both algorithms involve smoothing (or low-pass filtering) of the information signals. It is important to note, that when band-pass filtering is applied, the excellent time resolution of the SEOSA is blurred (and the event-detector property is somewhat lost). In brief, the effect of the Gabor filter is to smooth the spikes and the abrupt jumps (if any) of the original estimates (especially for the EOSA where high frequency components are preserved). Fig. 9

In real speech experiments we observe similar effects from the Gabor band-pass filter. The shape of the information signal estimates, though, is different from the linear synthetic case: The envelope and the instantaneous frequency are in many cases heavily modulated (especially for higher formants). In Fig. 9 we present a real speech example of resonance demodulation for the vowel /e/, around a formant (with center frequency) at $f_c \approx 3400$ Hz. A Gabor filter with center frequency at 3400 Hz and bandwidth parameter $\alpha = 1000$ Hz was used to extract the resonance. The HTSA and SEOSA estimated envelope are shown in (c), (e) respectively. The estimates present only minor differences at envelope minima and apart from that are almost identical. The instantaneous frequency estimates (d), (f) look also very similar. Note that the very small ripples that appear at the HTSA frequency estimate can be eliminated by increasing the length of the FIR filter that implements the discrete Hilbert transformer.

In numerous examples of speech analysis using SEOSA and HTSA, we saw only minor differences in the estimated amplitude envelope and instantaneous frequency contours. In some cases, the Hilbert transform algorithm seems to yield slightly smoother estimates than the SEOSA (especially

in frequency estimation and for lower formants). Also, in a few isolated instances, the SEOSA may produce narrow spikes (e.g. at envelope minima and at the corresponding places of the instantaneous frequency estimate). Note that the minor differences between the EOSA and HTSA estimators usually occur around the envelope minima. Overall it seems that both algorithms yield similar and equally satisfying results for real speech analysis. However, the SEOSA is faster and uses an extremely short analysis window.

4.2.2 On determining the Gabor filter parameters

A question that arises in speech demodulation experiments, is what happens to the envelope and the instantaneous frequency estimates, when the center frequency of the Gabor filter is not exactly equal to the center frequency of the formant. Using AM–FM speech-like signals, we observed that the estimated amplitude envelope and instantaneous frequency are close to the actual ones, for center frequency differences less than 100 Hz (and for sufficiently large filter bandwidth). In real speech experiments, shifting the center frequency of the Gabor filter (in the neighborhood of a formant) affects the envelope and frequency contours mainly around envelope minima. Specifically, the instantaneous frequency seems to be unstable around these points, presenting large peaks or valleys. In order to avoid such instabilities, the center formant frequency must be determined accurately. In [7], an iterative algorithm is proposed for determining the center formant frequency as the short-time average of the instantaneous frequency.

Determining the bandwidth of the Gabor filter is more difficult; an obvious choice is the bandwidth of the signal in question. Carson's rule implies that the bandwidth of an AM–FM signal is twice the sum of the maximum frequency deviation and the bandwidths of the AM and FM information signals. For our experiments, this would correspond to a bandwidth parameter α in the range of 3000-5000 Hz. In real speech, we need to isolate (via filtering) spectral peaks that are 500–1000 Hz apart. Thus, in order to avoid the effects of the neighboring formants (see section 4.3), we must limit the Gabor bandwidth to more conservative values, e.g. $\alpha = 1000$ Hz. Next, we consider how the envelope and instantaneous frequency estimates contours are affected when the bandwidth of the filter is smaller than the effective bandwidth of the signal.

We used synthetic AM–FM speech-like signals to address this question. The frequency modulation depth Ω_m/Ω_c parameter was selected to be 10%, (in speech analysis rarely have we found larger amounts of FM). For various values of AM, we found that a bandwidth parameter in the range $\alpha = 1000$ to 1500 Hz gives good envelope and frequency estimates (correct shape and maximum absolute error from 5%-10%).

In real speech, it is hard to determine how the bandwidth parameter α affects the envelope and instantaneous frequency, because of the effects of the neighboring formants. It is clear though, that for smaller α the bandwidth of the estimated information signals decreases.

Finally, there are algorithms for recovering the true information signals, by removing the blurring caused by the Gabor band-pass filter. For example, preliminary experiments have shown that the actual information signals $a(t)$ and $\omega_i(t)$ can be restored from the filtered amplitude and frequency estimates (40), (41) through deconvolution (for an accurately known carrier frequency).

4.3 Effects of Neighboring Spectral Peaks

A neighboring spectral peak that has not been thoroughly eliminated through band-pass filtering can seriously affect the estimated envelope and instantaneous frequency contours. In [11] a model has been proposed for dealing with this problem. Specifically, suppose that ω_c and ω_x are the center frequencies of a formant and its neighbor. Then the band-pass filtered signal $y(t)$ will be for $\lambda \ll 1$ (where λ is the relative gain of the neighboring formant vs. the center formant)

$$\begin{aligned} y(t) &= \cos(\omega_c t) + \lambda \cos(\omega_x t + \theta) \\ &\approx \cos[\omega_c t - \lambda \sin(\omega_f t - \theta)] + \lambda \cos(\omega_f t - \theta) \cos(\omega_c t) \end{aligned} \tag{42}$$

where $\omega_f = \omega_c - \omega_x$. Thus, the neighboring spectral peak modulates the envelope and instantaneous frequency estimates, with a modulation frequency ω_f equal to the difference of the central formant frequencies of the two spectral peaks.

Fig. 10

In Fig. 10 we present experimental evidence in support of this model. A synthetic speech vowel with formants at 550, 1550 and 2500 Hz and pitch frequency of 100 Hz is analyzed around $f_c = 2500$ Hz. We have seen in Fig. 8, how the estimated envelope and frequency curves look when the neighboring formants have been thoroughly eliminated through Gabor filtering with the appropriate bandwidth. In our example we choose the bandwidth parameter to be $\alpha = 1550$ Hz so that the formant peak at 1550 Hz is still in play. The estimates for the envelope and the instantaneous frequency are displayed in Fig. 10 (a) and (b). The estimates are clearly modulated, with a modulation frequency equal to the difference between the two formant frequencies, i.e. 950 Hz. The amplitude of the modulations increases as the Gabor filter bandwidth increases. We have observed similar phenomena in real speech, due to neighboring formants.

4.4 Analysis of Transitions in Real Speech

Formant finding and feature extraction during speech transitions is not a simple task, as most speech parameters change rapidly and short-time analysis assumptions do not hold. The usual approach to this problem is to use continuity and smoothness constraints to interpolate the data from neighboring frames. Unfortunately, this method does not always produce satisfactory results.

We know that the energy separation algorithm has better time resolution than conventional short-time analysis methods (the estimates are computed at each sample instead of once over each analysis frame). Also, the Gabor filter followed by the SEOSA introduces minimal blurring/smoothing of rapidly varying speech modulation features. Thus, one can use the amplitude envelope and instantaneous frequency estimates for parameter estimation and feature extraction during transitions in real speech. Fig. 11

An example of tracking a transient formant frequency is presented in Fig. 11, during a transition from the unvoiced consonant /c/ to the vowel /e/. In our example, the vowel /e/ has a strong spectral peak around 4500 Hz, while the consonant /c/ does not have a formant around this frequency. We band-pass filter the speech signal, using a Gabor filter with center frequency at $f_c = 4850$ Hz and bandwidth parameter $\alpha = 1500$ Hz, in order to follow the formant track during the transition from /c/ to /e/. The SEOSA absolute envelope estimate (c) shows the energy increase as we pass from the unvoiced to the voiced sound. Similar amplitude modulation patterns can be seen in both sounds. The instantaneous frequency estimate (d) shows clearly the center frequency of the formant changing rapidly during the transition. We observe a 500 Hz shift in the center formant frequency in a time period of approximately 5-10 msec. Note also that the vowel reaches a steady state 10-15 msec after voicing begins.

4.5 An Application: Formant Tracking

In this section we describe an iterative energy separation algorithm [7] (iterative SEOSA) for tracking the center formant frequencies of a speech signal. Once the amplitude envelope $a(t)$ and instantaneous frequency $\omega_i(t)$ of a formant are known (via the SEOSA), there are various ways of estimating the center formant frequency (over a short time interval). The simpler estimate is ω_{AF}, the average value of the instantaneous frequency $\omega_i(t)$

$$\omega_{AF} = \frac{1}{T} \int_{t_0}^{t_0+T} \omega_i(t)\, dt \tag{43}$$

where t_0 is the start and T the duration of the speech analysis window. It can be seen through a crude voiced speech model (each formant band is a sum of harmonics), that the estimate ω_{AF} can

lead us to the center formant frequency through an iterative procedure. According to this model, at local envelope maxima the instantaneous frequency accurately tracks the formant value, while at local envelope minima the frequency curves present spikes that point towards the speech harmonic with the higher amplitude (in the bandpassed speech spectrum). Thus, if we iteratively use ω_{AF} as the center frequency of the Gabor filter and perform the SEOSA on the extracted formant band, the refined estimate ω_{AF} will converge to the formant frequency after a few iterations.

We claim above, that at envelope maxima the instantaneous frequency takes values that correspond to what we intuitively consider to be the center formant frequency at that instant. As an example, consider the envelope and instantaneous frequency of the sum of two sinusoids with time varying amplitudes $a_1(t)$, $a_2(t)$ and constant frequencies ω_1, ω_2. The instantaneous frequency at the instants t_n where local envelope maxima occur can be seen to be:

$$\omega_i(t_n) = \frac{a_1(t_n)\,\omega_1 + a_2(t_n)\,\omega_2}{a_1(t_n) + a_2(t_n)} \tag{44}$$

Keeping this result in mind, we can refine further the center formant frequency estimate (after the above described iterative procedure has reached a formant peak) by computing the average instantaneous frequency at time intervals around the local envelope maxima only.

Another formant frequency estimate [6, 1] that has been used in the past for formant tracking [3] is the first central moment of the spectrum of the signal:

$$<\omega> = \frac{\int_{-\infty}^{+\infty} \omega\,|Z(\omega)|^2\,d\omega}{\int_{-\infty}^{+\infty} |Z(\omega)|^2\,d\omega} \tag{45}$$

where $Z(\omega)$ is the Fourier transform of the analytic signal $z(t)$ (8). It can be shown [27, 10] that the average frequency $<\omega>$ in the spectrum is equal to the weighted time average $<\omega_i>$ of the instantaneous frequency defined as:

$$<\omega_i> = \frac{\int_{-\infty}^{+\infty} \omega_i(t)\,|z(t)|^2\,dt}{\int_{-\infty}^{+\infty} |z(t)|^2\,dt} \tag{46}$$

where $z(t)$ is the analytic signal (8). We saw in Section 2.1, that for small quadrature error $e(t)$, the modulus $|z(t)|$ of the analytic signal is an accurate estimate of the amplitude envelope $|a(t)|$. Thus, for small quadrature error, the above estimate of the center formant frequency can be expressed as a function of the envelope $a(t)$ and instantaneous frequency $\omega_i(t)$ as:

$$\omega_{WAF} = \frac{\int_{t_0}^{t_0+T} \omega_i(t)\,a(t)^2\,dt}{\int_{t_0}^{t_0+T} a(t)^2\,dt} \tag{47}$$

where T is the duration of the analysis window. The estimate can be refined iteratively in a scheme similar to the one described above. Attention must be paid to the boundary conditions when using

the short-time equation (47); for frames 2-3 pitch period long, though, the boundary effects are negligible.

In order to start the iterative SEOSA, we must first find some initial formant estimates for each analysis frame. Raw formant estimates can be obtained either from the roots of the LPC polynomial or from morphological peak picking of the speech spectrum [7]. Next, the speech signal is filtered through a bank of Gabor bandpass filters with center frequencies equal to the raw formant estimates and the SEOSA is applied on each speech band. The average instantaneous frequency ω_{AF} or the weighted average frequency ω_{WAF} is computed for each formant band and is used as the center frequency of the Gabor filter for the next step of the iteration. The bandwidth of the filter is constant (typically the bandwidth parameter α is 1000 Hz). This procedure is repeated until convergence is reached; then the next analysis frame is processed (short-time analysis). Fig. 12

In Fig. 12 an example of the iterative formant tracking algorithm is displayed for the word 'thevenin' (a). The center formant frequency values, where the iterative formant tracking algorithm converged, are shown for consecutive analysis frames (20 msec of duration, 300 samples) at (b),(c). The center formant frequency estimate is ω_{AF} and ω_{WAF} for (b) and (c) respectively. Finally, the LPC raw formant tracks are shown (d), obtained from peak-picking of the LPC smooth spectrum (LPC order is 20, sampling frequency for 'thevenin' is 15 KHz). All three algorithms perform well. The iterative algorithm (b), (c) though, provides more detailed formant tracks than the LPC formant tracker (d) (e.g. observe the second and fifth formant track). Note also that the first formant estimates are more accurate when ω_{AF} is used. Detailed comparisons between the two formant tracking algorithms based on the iterative SEOSA are currently being performed.

Finally, an alternative formant tracking algorithm using the ω_{AF} and ω_{WAF} estimates is the parallel multiband implementation, where the signal is filetered through a bank of Gabor bandpass filters with constant center frequencies that cover the spectrum range of interest. The number of filters may vary: in our implementation we have used filters with center frequencies 50-100 Hz apart. For each speech band the center formant frequency estimate (ω_{AF} or ω_{WAF}) is calculated. When the values of the formant estimates $\omega_{AF,\,WAF}$ of two neighboring bands lie in the interval bounded by the center frequencies of the bands (i.e. the Gabor filter center frequencies), we conclude that a formant exists in that interval. The exact formant value is calculated by linear interpolation of the formant estimates of the two adjacent bands. This algorithm needs no initial estimates and provides good formant tracks.

5 Conclusions

In this paper, we have compared two different approaches for estimating the time-varying amplitude envelope and instantaneous frequency of general AM–FM signals, as well as of speech resonances: the energy operator separation algorithm (EOSA) involving a nonlinear differential operator and the Hilbert transform separation algorithm (HTSA) involving a linear integral transform. We have also proposed a refinement of the EOSA, called the smoothed energy operator separation algorithm (SEOSA), that implements a very short binomial FIR smoothing of the energy signals. We have compared the three algorithms first on general synthetic AM–FM signals and then on speech vowel resonance signals extracted via Gabor band-pass filtering. Next, some important issues related to the application of the separation algorithms to speech resonance demodulation were investigated and briefly discussed. These include choosing the Gabor band-pass filter parameters, the effect of neighboring formants, and transient formant analysis. Finally, an application of the SEOSA to formant tracking was presented.

After extensive experiments on synthetic AM–FM signals, we have found that, in the absence of noise, when the ratio R of the carrier frequency vs. the information signals' bandwidth is in the order of 10 (as in speech applications) the EOSA yields a mean absolute error in the order of $10^{-1}\%$; when this ratio becomes 100 or 1000 (as in communication applications) the EOSA yields errors in the order of $10^{-2}\%$ and $10^{-3}\%$, respectively. Note that even in the worst case ($R = 10$) the EOSA yields a relatively small error. The SEOSA almost always reduces the EOSA error by about 50%, except for very high values of R. The HTSA yields an error in the order of $10^{-2}\%$ for all the above values of R. In the presence of 30 db noise, both HTSA ans SEOSA yield errors in the order of 1%, with the SEOSA yielding the smallest error. In the analysis of short time segments of speech vowel signals (synthetic or real) band-pass filtered around their formants, the SEOSA was found to yield modulating signals very close to the ones obtained via the HTSA. The fact that both algorithms yield similar results for speech signals is due to the band-pass filtering, which blurs the instantaneously varying features of the time waveform.

For all signals, both EOSA and SEOSA have very small computational complexity, linear in the number of input samples. The HTSA has about one order of magnitude higher complexity. For speech applications the HTSA complexity becomes quadratic in the number of samples. Finally, while the EOSA or SEOSA requires for its operation an extremely small number of input samples in its moving window, the HTSA requires an order of magnitude longer window.

In short, the SEOSA was found to yield comparable estimation errors to the HTSA for tracking

AM–FM modulations in speech signals. For communications applications, the SEOSA yields a smaller error. In addition, the SEOSA has the advantages over the HTSA of smaller computational complexity and faster time-adaptivity. Finally, the use of an energy operator gives the SEOSA an additional interesting intuitive feature: the tracking of the energy required for generating the AM–FM signal and the separation of the energy into amplitude and frequency components.

Acknowledgements

We wish to thank Alan Bovik at the University of Texas at Austin, Jim Kaiser at Rutgers University, Tom Quatieri at MIT Lincoln Laboratory, and Ron Schafer at Georgia Institute of Technology for many insightful discussions on various topics presented in this paper.

References

[1] B. Boashash, " Estimating and Interpreting the Instantaneous Frequency of a Signal – Part 1: Fundamentals", Proceedings of the IEEE, vol. 80, no. 4, pp. 520-538, Apr. 1992.

[2] A. C. Bovik, P. Maragos, and T. F. Quatieri, "Measuring Amplitude and Frequency Modulations in Noise Using Multiband Energy Operators", to appear in IEEE Int'l Symp. on Time-Frequency and Time-Scale Analysis, Victoria, B. C., Canada, Oct. 1992.

[3] J. L. Flanagan, Speech Analysis Synthesis and Perception, 2nd edition, Springer-Verlag, 1972.

[4] J. L. Flanagan, "Parametric Coding of Speech Spectra", J. Acoust. Soc. Am., vol. 68, pp. 412-419, Aug. 1980.

[5] J. Foote, D. Mashao and H. Silverman, "Stop Classification Using DESA–1 High–Resolution Formant Tracking", in Proc. IEEE Int'l Conf. Acoust., Speech and Signal Processing, Minneapolis, MN, Apr. 1993.

[6] D. Gabor, "Theory of Communication", J. IEE (London), vol. 93, pp. 429-457, 1946.

[7] H. M. Hanson, P. Maragos, A. Potamianos, "Finding Speech Formants and Modulations via Energy Separation: With Application to a Vocoder", in Proc. IEEE Int'l Conf. Acoust., Speech and Signal Processing, Minneapolis, MN, Apr. 1993.

[8] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", in Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Processing, Albuquerque, New Mexico, pp. 381–384, Apr. 1990.

[9] J. F. Kaiser, "On Teager's Energy Algorithm and Its Generalization to Continuous Signals", in Proc. IEEE Digital Signal Processing Workshop, New Paltz, NY, Sep. 1990.

[10] L. Mandel, "Interpretation of instantaneous frequency", Am. J. Phys., vol. 42, pp. 840-846, 1974.

[11] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", Tech. Report 91–17, Harvard Robotics Lab., Harvard Univ., Nov. 1991. Also submitted to IEEE Transactions on Signal Processing.

[12] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On Separating Amplitude from Frequency Modulations Using Energy Operators", in Proc. IEEE Int'l Conf. Acoust., Speech, and Signal Processing, San Francisco, CA, Mar. 1992.

[13] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On Amplitude and Frequency Demodulation Using Energy Operators", IEEE Transactions on Signal Processing, vol. 41, no 4, pp. 1532-1550, Apr. 1993.

[14] P. Maragos, T. F. Quatieri and J. F. Kaiser, "Detecting Nonlinearities in Speech using an Energy Operator", in Proc. IEEE Digital Signal Processing Workshop, Mohonk (New Paltz), NY, Sep. 1990.

[15] P. Maragos, T. F. Quatieri, J. F. Kaiser, "Speech Nonlinearities, Modulations, and Energy Operators", in Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Processing, Toronto, Ontario, Canada, May 1991.

[16] A. H. Nuttall, "On the Quadrature Approximation to the Hilbert Transform of Modulated Signals", Proc. IEEE, vol. 54, pp. 1458-1459, 1966.

[17] A. H. Nuttall, "Complex Envelope Properties, Interpretation, Filtering and Evaluation", Tech. Report TR 8827, Naval Underwater Systems Center, Feb. 1991.

[18] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, Prentice Hall, NJ, 1989.

[19] A. Papoulis, The Fourier Transform and Its Applications, McGraw-Hill, NY, 1962.

[20] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, NY, 1984.

[21] T. F. Quatieri, R. B. Dunn, P. Maragos, J. F. Kaiser, and A. C. Bovik, "Detection of Transient Signals Using the Energy Operator", Tech. Report, MIT Lincoln Lab., in preparation.

[22] T. F. Quatieri, J. F. Kaiser and P. Maragos, "Transient Detection in AM–FM Background using an Energy Operator", presented at the 1991 IEEE Underwater Acoustic Signal Processing Workshop, Univ. of Rhode Island, Oct. 1991.

[23] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice Hall, NJ, 1978.

[24] M. Schwartz, Information Transmission, Modulation and Noise, McGraw-Hill, New York, NY, 1980.

[25] H. M. Teager, "Some Observations on Oral Air Flow During Phonation", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP–28, pp. 599–601, Oct. 1980.

[26] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", NATO Advanced Study Institute on Speech Production and Speech Modeling, Bonas, France, July 1989; Kluwer Acad. Publ., Boston, MA, pp. 241–261, 1990.

[27] J. Ville, " Theory et applications de la notion de signal analytique ", Cable et Transmission, vol. 2A, pp. 61-74, 1948.

[28] A. Zayezdny and I. Druckmann, "A new method of signal description and its applications to signal processing", Signal Processing, vol. 22, pp. 153–178, Feb. 1991.

Footnotes

(0) When this paper was first submitted the authors were at the Division of Applied Sciences, Harvard University, Cambridge, MA 02138, USA. They are currently at the School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

(1) This research work was supported by the National Science Foundation under Grant MIP-91-20624, by a NSF Presidential Young Investigator Award under Grant MIP-86-58150 with matching funds from Xerox, and by a Wallace Fellowship.

(2) Smoothing the output of the energy operator via low-pass filtering to reduce estimation errors has also been done in [21] when using the energy operator as detector of transient signal signatures in AM–FM background noise.

TABLE 1: Percent Amplitude and Frequency Estimation Mean Absolute Errors using Separation Algorithms on AM–FM/Cosine Signals ($\Omega_c = \pi/5$)

	$\Omega_c/\Omega_a = \Omega_c/\Omega_f = 10$		$\Omega_c/\Omega_a = \Omega_c/\Omega_f = 100$	
Algorithm	Amplitude error %	Frequency error %	Amplitude error %	Frequency error %
short HTSA	4.41	4.60	4.39	4.46
long HTSA	0.03	0.04	0.03	0.03
EOSA	0.39	0.32	0.03	0.02
smooth EOSA	0.25	0.22	0.02	0.01

TABLE 2: Percent Amplitude and Frequency Estimation Mean Absolute Errors using Separation Algorithms on AM–FM/Cosine Signals with 30 db Noise

	$\Omega_c/\Omega_a = \Omega_c/\Omega_f = 10$		$\Omega_c/\Omega_a = \Omega_c/\Omega_f = 100$	
Algorithm	Amplitude error %	Frequency error %	Amplitude error %	Frequency error %
short HTSA	4.87	5.06	4.85	4.93
long HTSA	1.83	2.15	1.84	2.19
EOSA	4.81	4.43	4.71	4.35
smooth EOSA	1.37	1.87	1.28	1.72

TABLE 3: Computational Complexity of Separation Algorithms.
(Number of Operations Per Sample)

Algorithm	Additions	Multiplications	$\arccos(\cdot)$	$\sqrt{(\cdot)}$	W
short HTSA	12	8	1	1	20
long HTSA	73	38	1	1	140
EOSA	6	8	1	1	5
smooth EOSA	24	8	1	1	11

* W is the number of samples in the moving window.

List of Figures

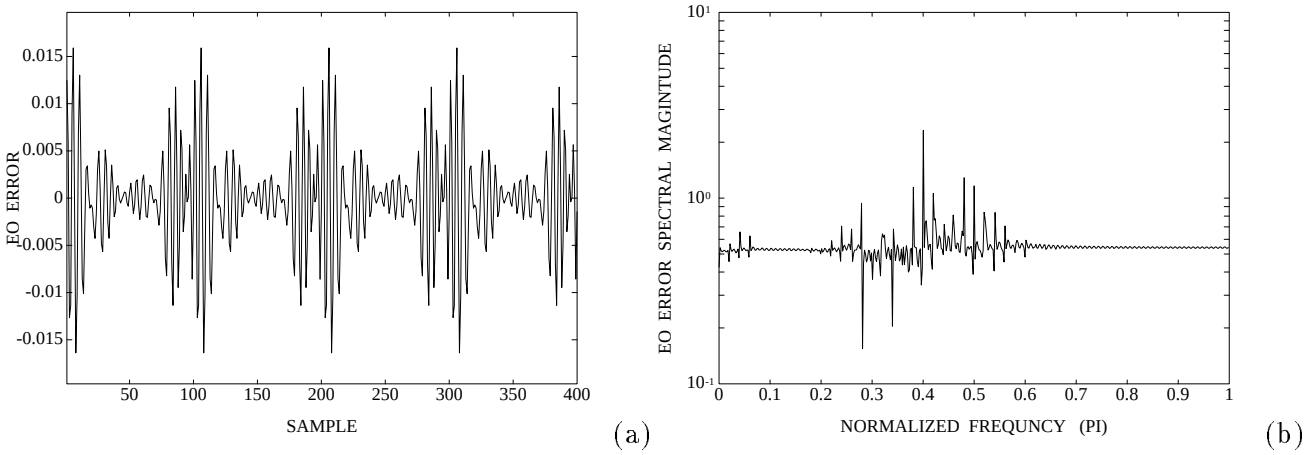


Figure 1: (a) The energy operator approximation error for the AM–FM signal $x(n) = (1 + 0.5\cos(\pi n/50))\cos[\pi n/5 + \sin(\pi n/25) + \phi]$. (b) The magnitude of the Fourier transform of the approximation error.

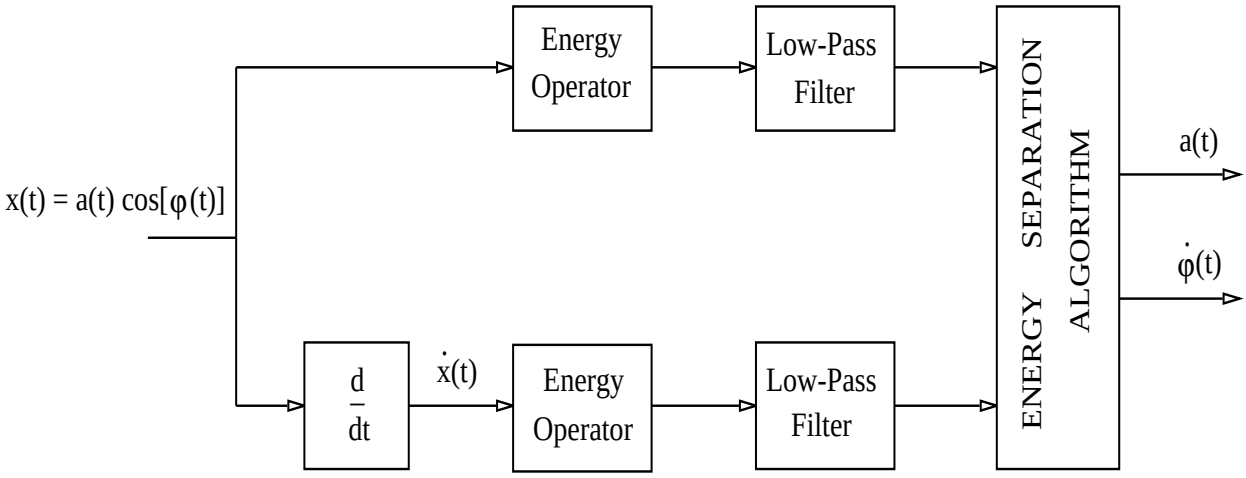


Figure 2: The block diagram of the Smoothed Energy Operator Separation Algorithm.

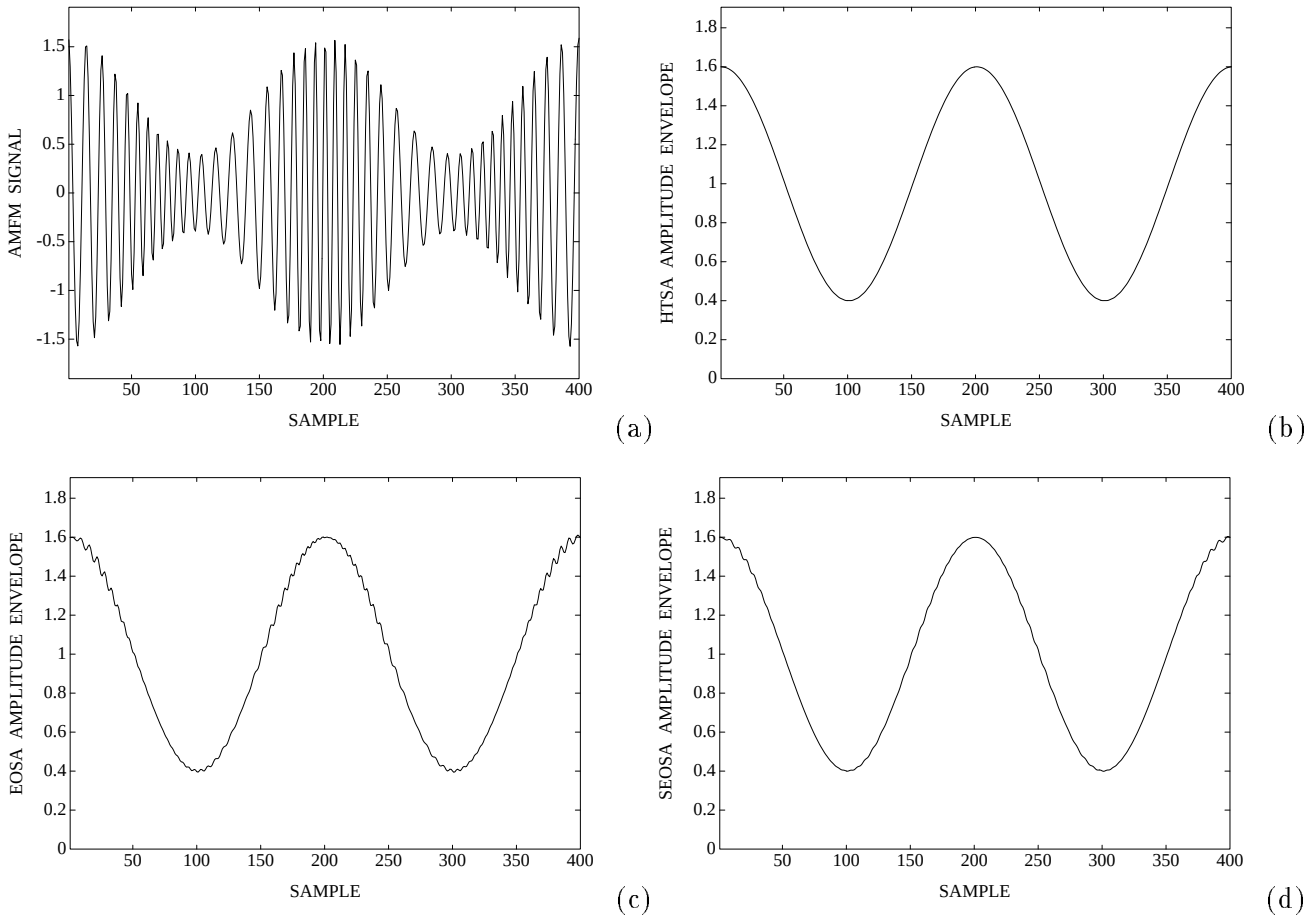


Figure 3: (a) AM–FM signal $x(n) = (1 + 0.6\cos(\pi n/100))\cos[\pi n/5 + 4\sin(3\pi n/200)]$. Estimated amplitude envelope using the: (b) Hilbert Transform Separation Algorithm (HTSA), (c) Energy Operator Separation Algorithm (EOSA), and (d) Smoothed Energy Operator Separation Algorithm (SEOSA).

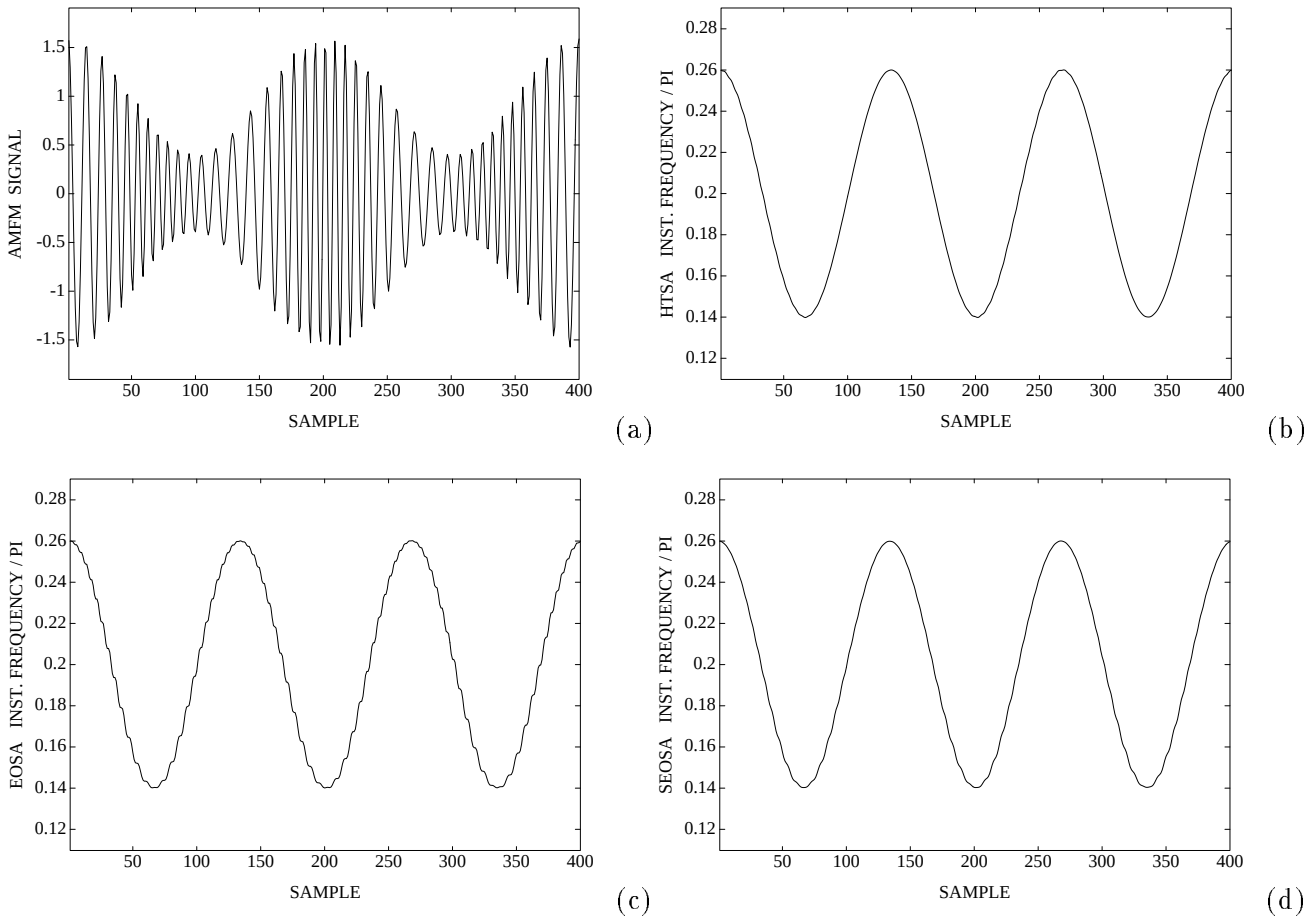
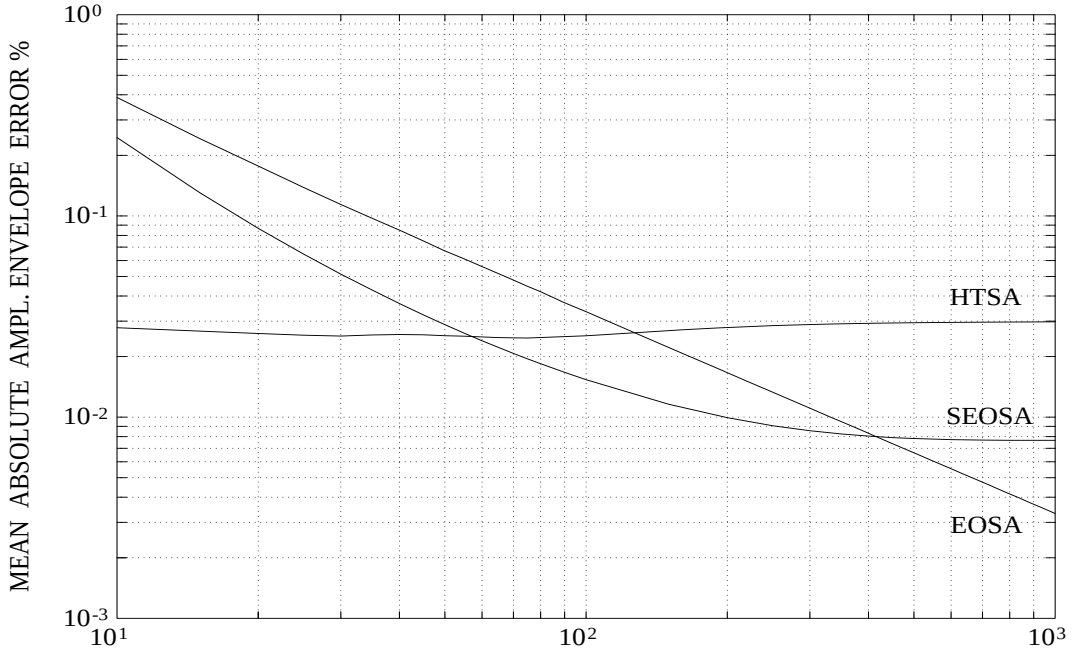
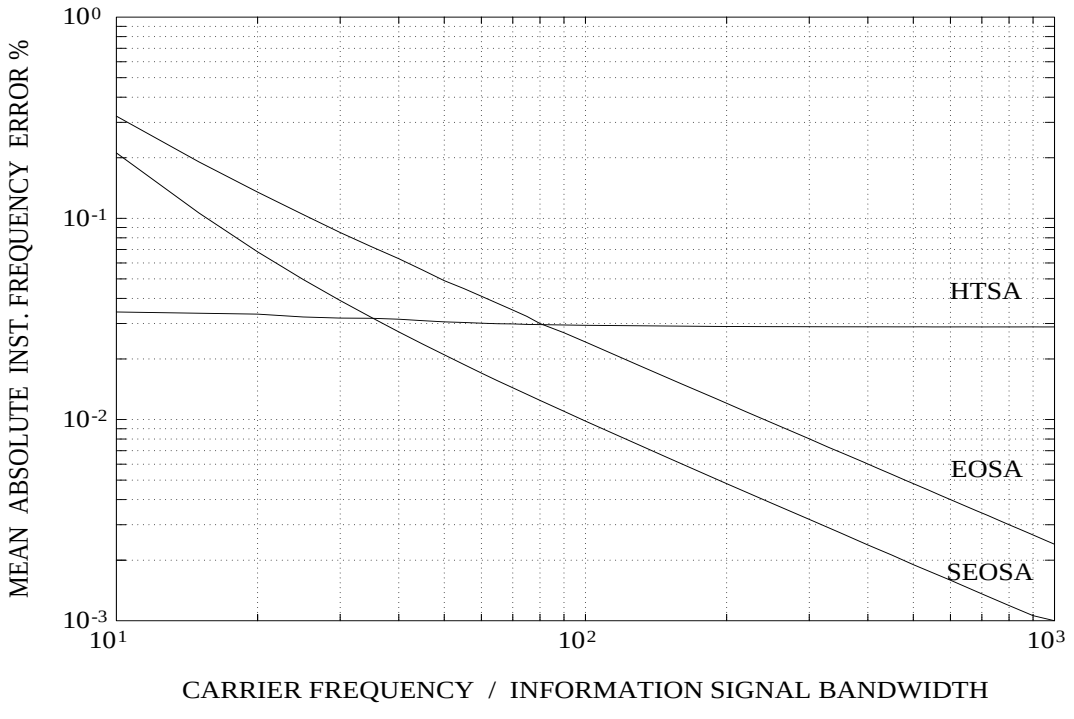


Figure 4: (a) AM–FM signal $x(n) = (1 + 0.6\cos(\pi n/100))\cos[\pi n/5 + 4\sin(3\pi n/200)]$. Estimated instantaneous frequency using the: (b) HTSA, (c) EOSA, and (d) SEOSA.



(a)



(b)

Figure 5: (a) EOSA, SEOSA and HTSA mean absolute envelope estimation error (%) for $\Omega_c/\Omega_{a,f} \in [10, 1000]$ ($\Omega_c = \pi/5$ is the carrier frequency and $\Omega_{a,f}$ is the bandwidth of the AM and FM modulating signals.) (b) EOSA, SEOSA and HTSA mean absolute frequency estimation error (%). Each point in these curves is the average error over 100 experiments on AM–FM/Cosine signals as the percent of AM varies from 5-50% and of FM from 2-20%.

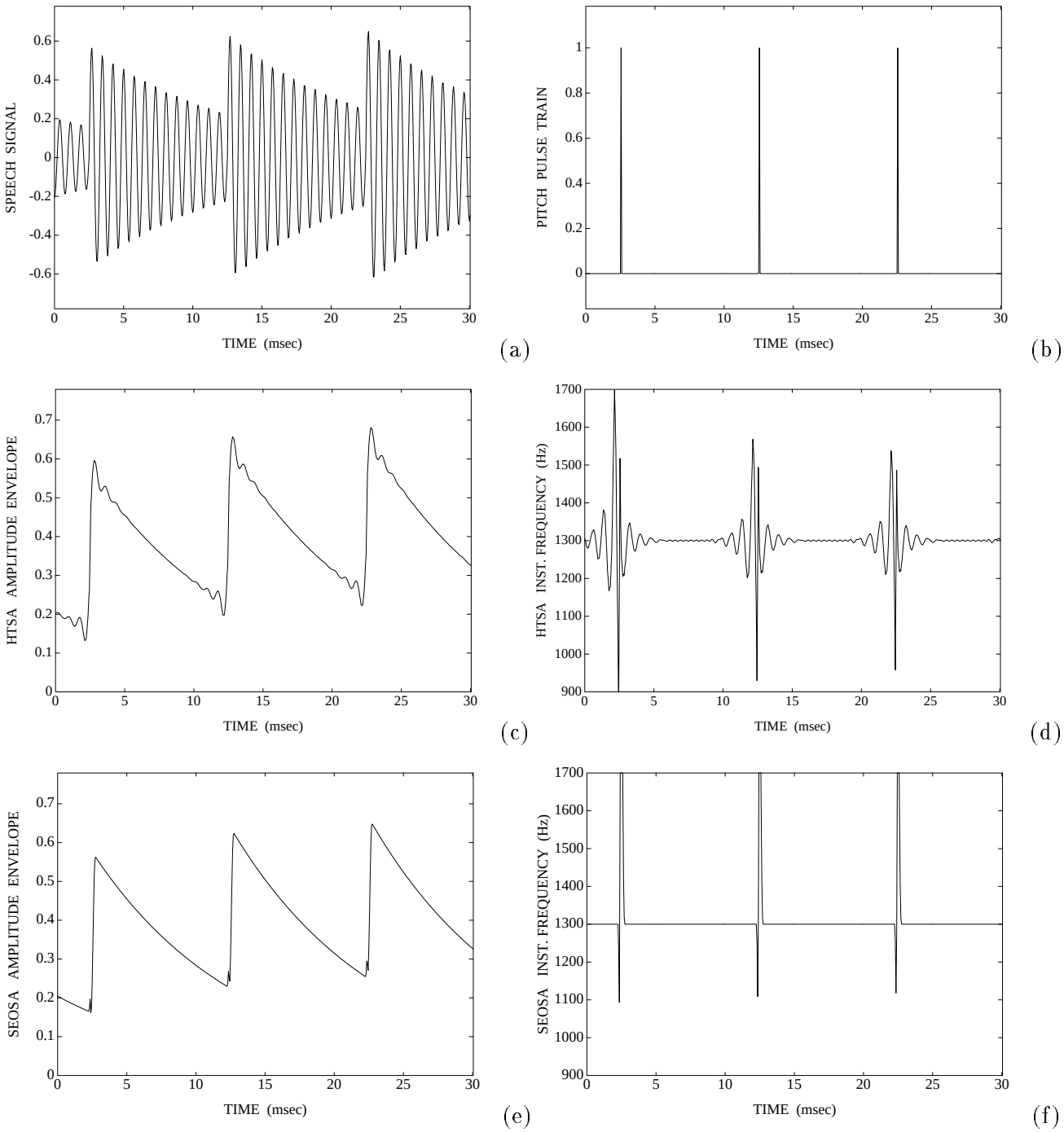


Figure 6: (a) Synthetic speech signal with a single formant at 1300 Hz and pitch frequency at 100 Hz (sampling frequency at 20 KHz). (b) The excitation, a sequence of pulses with a 10 msec period. (c) Estimated amplitude envelope using HTSA (d) Estimated instantaneous frequency using HTSA (e) Estimated amplitude envelope using SEOSA (f) Estimated instantaneous frequency using SEOSA

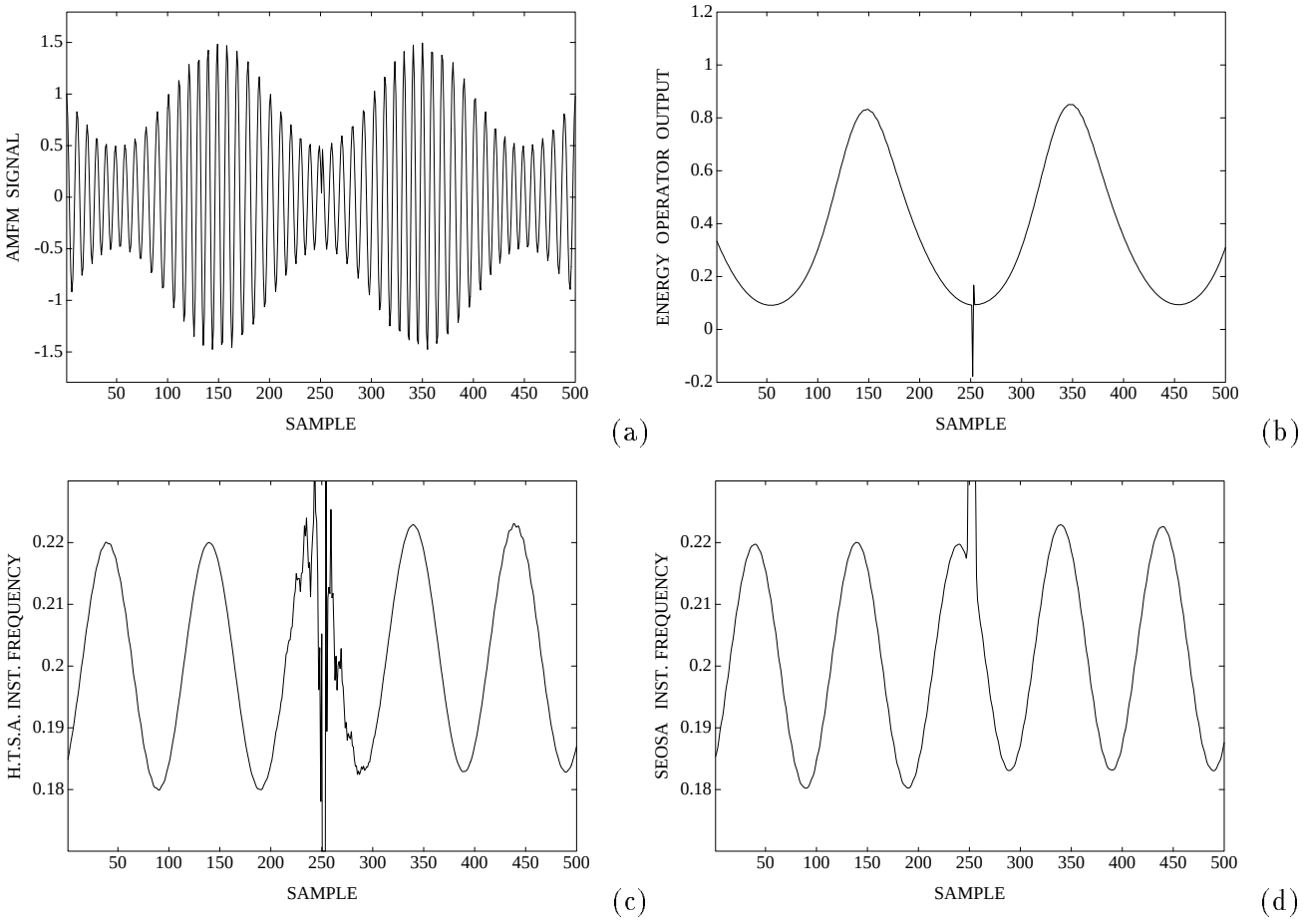


Figure 7: (a) The AM–FM signal $x(n) = (1 + 0.5\cos(\pi n/100))\cos[a\pi n/5 + 2\sin(\pi n/50) + \phi]$, where $a = 1$ for the first 250 samples and $a = 1.0143$ for the rest. (b) The Energy Operator output $\Psi[x(n)]$. Estimated instantaneous frequency of $x(n)$ using: (c) HTSA, (d) SEOSA.

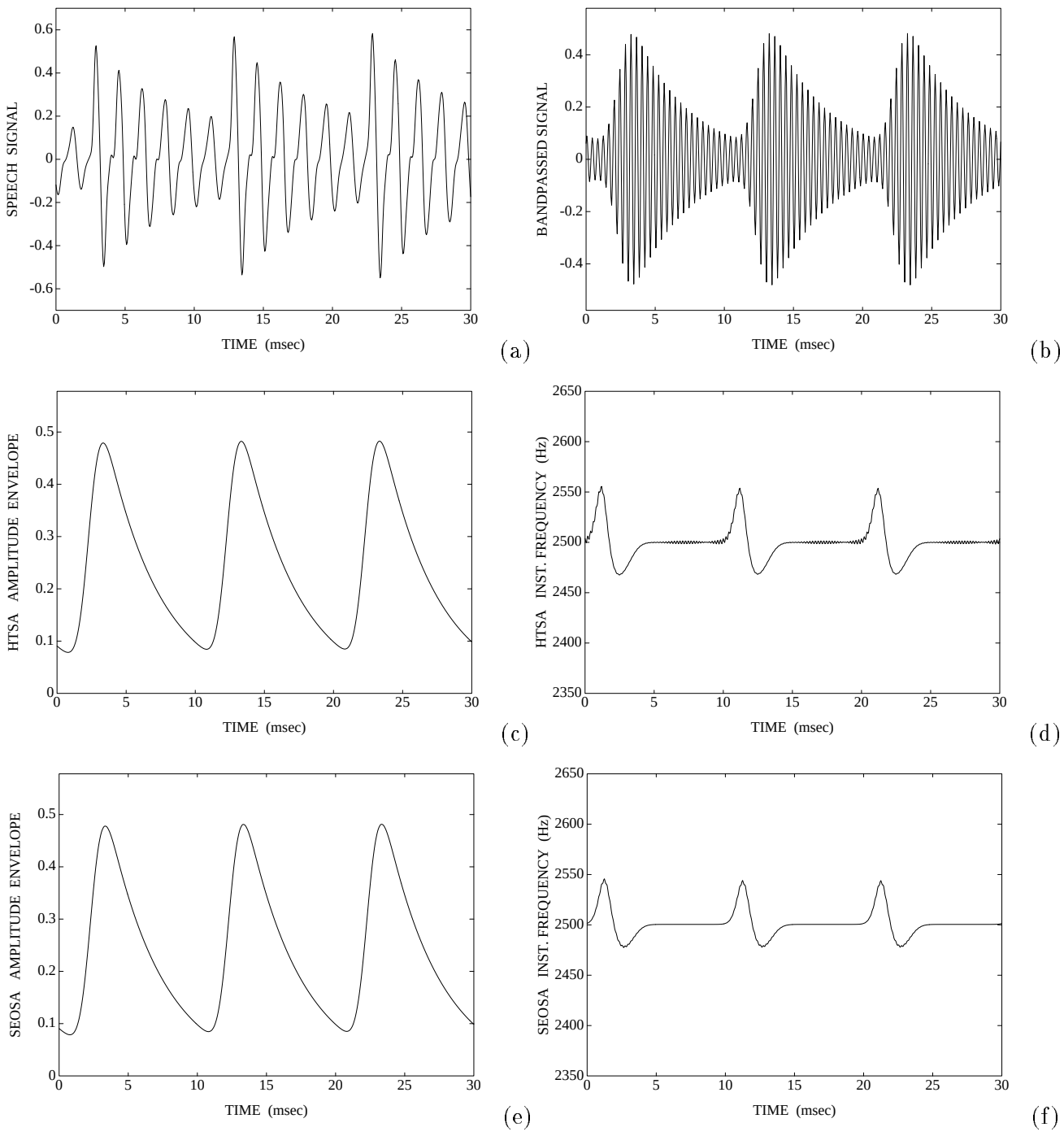


Figure 8: (a) Synthetic speech signal $x(n)$ (vowel /a/) with formants at 600, 1200, 2500 and 3600 Hz and pitch frequency at 100 Hz (sampling frequency at 20 KHz). (b) Speech signal after Gabor filtering around the formant at $f_c = 2500$ Hz (filter bandwidth parameter $\alpha = 1000$ Hz). (c) Estimated amplitude envelope using HTSA (d) Estimated instantaneous frequency using HTSA (e) Estimated amplitude envelope using SEOSA (f) Estimated instantaneous frequency using SEOSA

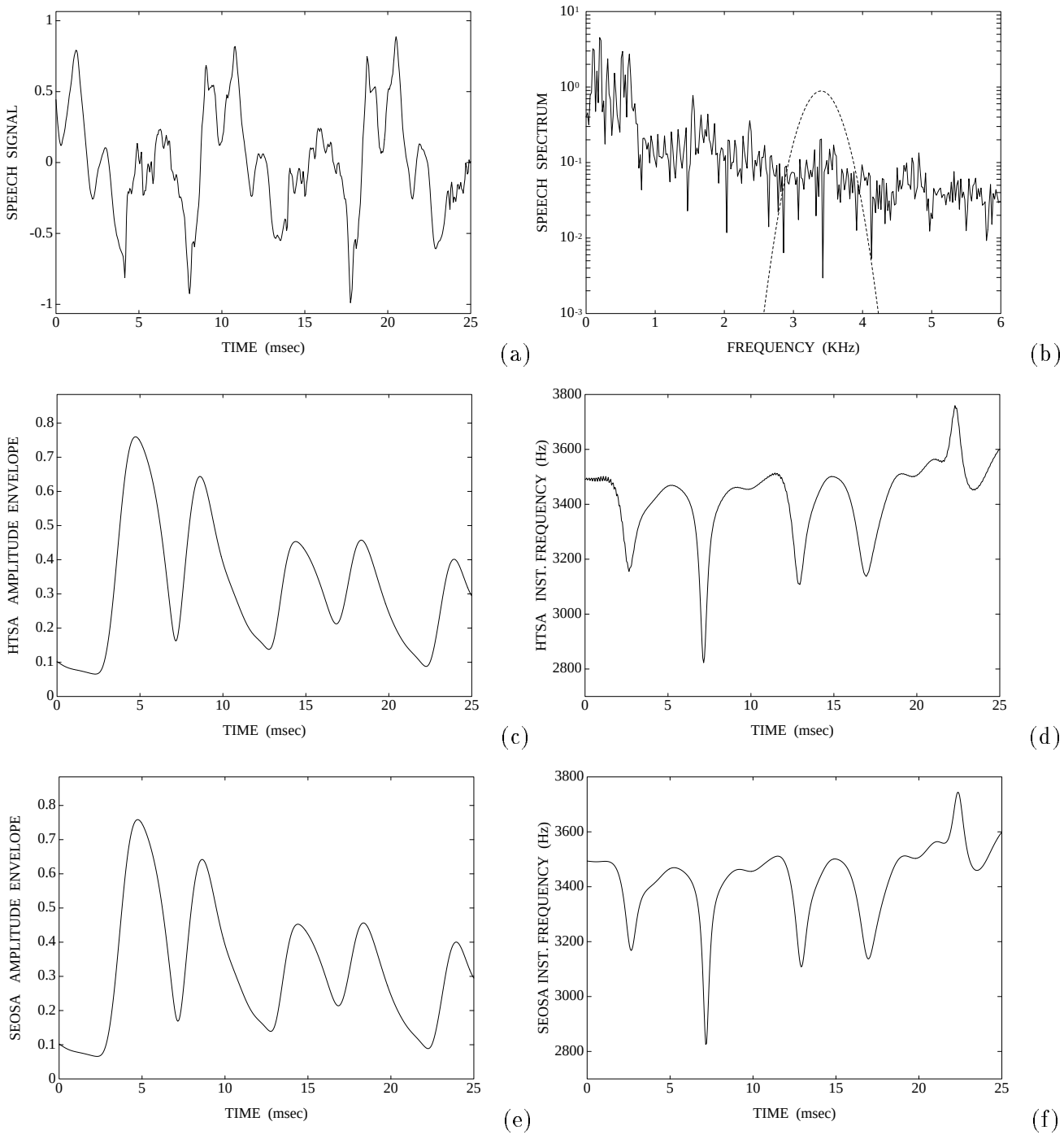


Figure 9: (a) Speech signal $s(n)$ (vowel /e/ sampled at 20 KHz). (b) Spectral magnitude of the speech signal $s(n)$ and of the Gabor filter ($f_c = 3400$ Hz, $\alpha = 1000$ Hz). (c) Estimated amplitude envelope using HTSA, after Gabor filtering of $s(n)$ around the spectral peak at $f_c = 3400$ Hz. (d) Estimated instantaneous frequency using HTSA (e) Estimated amplitude envelope using SEOSA (f) Estimated instantaneous frequency using SEOSA

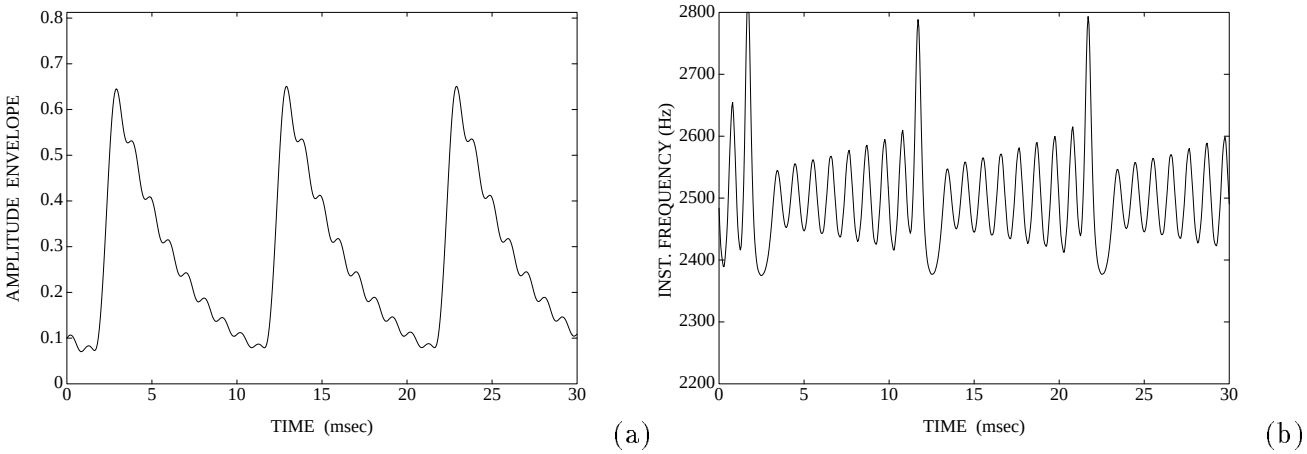


Figure 10: (a) SEOSA estimated amplitude envelope for the third formant ($f_c = 2500$ Hz) of the band-passed synthetic speech vowel $x(n)$ (formants at 550, 1550, 2500 Hz and pitch at 100 Hz, Gabor center frequency at $f_c = 2500$ Hz and bandwidth parameter $\alpha = 1550$ Hz). (b) Estimated instantaneous frequency using EOSA.

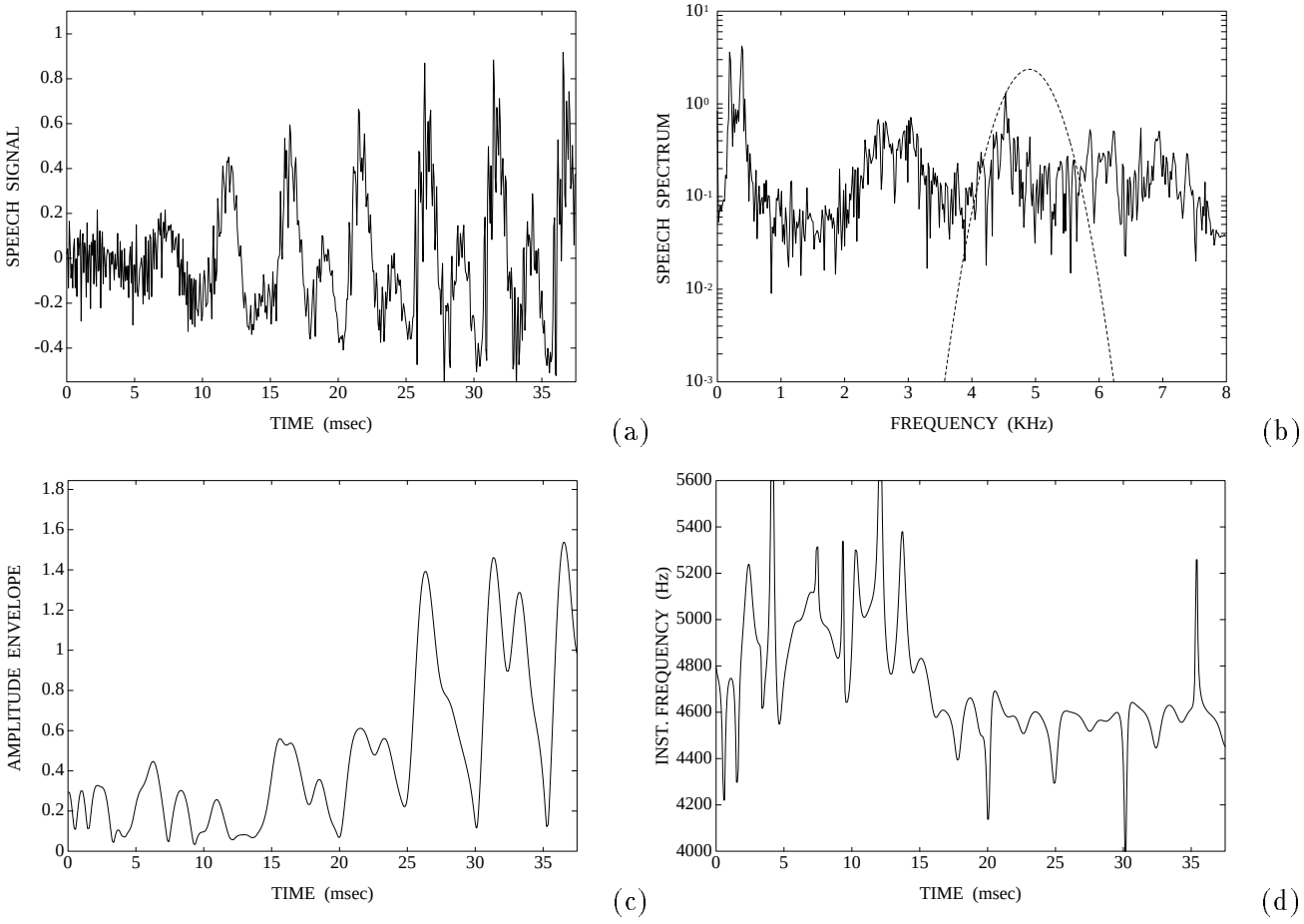


Figure 11: (a) Speech signal $x(n)$ (transition between unvoiced consonant /c/ and vowel /e/ sampled at 16 KHz). (b) Spectral magnitude of the speech signal $x(n)$ and of the Gabor filter ($f_c = 4850$, $\alpha = 1500$ Hz). (c) Estimated amplitude envelope using SEOSA, after Gabor filtering of $x(n)$ around $f_c = 4850$ Hz ($\alpha = 1500$ Hz). (d) Estimated instantaneous frequency using SEOSA

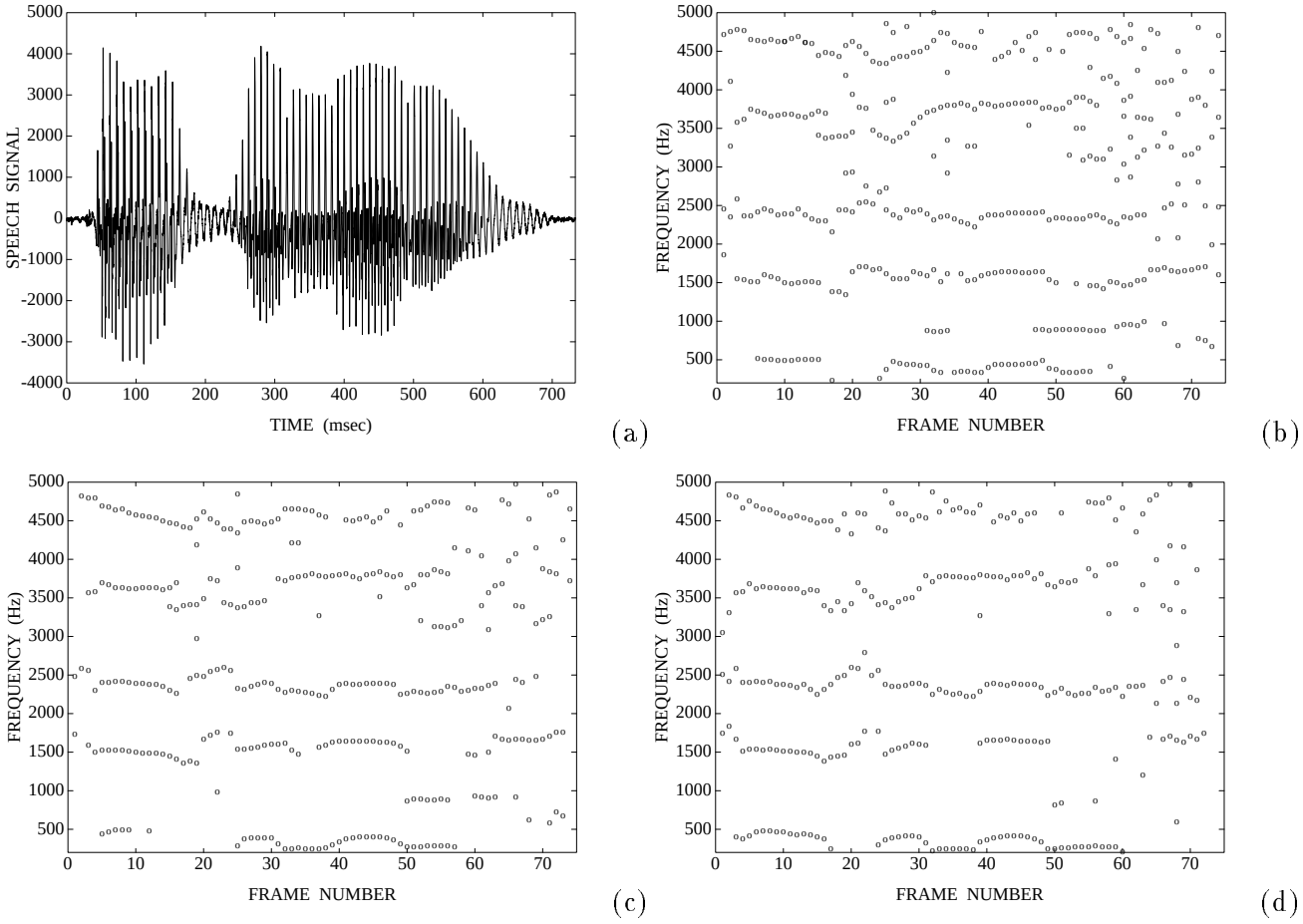


Figure 12: (a) The word 'thevenin' (sampled at 15 KHz). (b) Formant tracks for 'thevenin' using the iterative SEOSA and the average instantaneous frequency ω_{AF} as the center formant frequency estimate (frame duration 20 msec, updated every 10 msec). (c) Formant tracks using the iterative SEOSA and the weighted average instantaneous frequency ω_{WAF}. (d) Formant tracks from peak-picking of the LPC spectrum (LPC order is 20).