Energy Separation in Signal Modulations with Application to Speech Analysis

Petros Maragos, *Senior Member, IEEE*, James F. Kaiser, *Fellow, IEEE*, and Thomas F. Quatieri, *Senior Member, IEEE*

Abstract—**Oscillatory signals that have both an amplitude-modulation (AM) and a frequency-modulation (FM) structure are encountered in almost all communication systems. We have also used these structures recently for modeling speech resonances, being motivated by previous work on investigating fluid dynamics phenomena during speech production that provide evidence for the existence of modulations in speech signals. In this paper, we use a nonlinear differential operator that can detect modulations in AM–FM signals by estimating the product of their time-varying amplitude and frequency. This operator essentially tracks the energy needed by a source to produce the oscillatory signal. To solve the fundamental problem of estimating both the amplitude envelope and instantaneous frequency of an AM–FM signal we develop a novel approach that uses nonlinear combinations of instantaneous signal outputs from the energy operator to separate its output energy product into its amplitude modulation and frequency modulation components. The theoretical analysis is done first for continuous-time signals. Then several efficient algorithms are developed and compared for estimating the amplitude envelope and instantaneous frequency of discrete-time AM–FM signals. These energy separation algorithms are then applied to search for modulations in speech resonances, which we model using AM–FM signals to account for time-varying amplitude envelopes and instantaneous frequencies. Our experimental results provide evidence that bandpass filtered speech signals around speech formants contain amplitude and frequency modulations within a pitch period. Overall, the energy separation algorithms, due to their very low computational complexity and instantaneously-adapting nature, are very useful in detecting modulation patterns in speech and other time-varying signals.**

I. Introduction

MODULATION of the amplitude and/or frequency of a sine wave has been used extensively in communication systems for transmitting information. Some of the early pioneering work in this field was done by Carson [3], van der Pol [37], [38], and Armstrong [1]. The recent state of basic knowledge on amplitude-modulation (AM) and frequency-modulation (FM) systems can be found in contemporary books on communications, e.g., [27]. Recently [18], we have also used AM and FM models to represent time-varying amplitude and frequency patterns in speech resonances. Our initial inspiration for attempting to model and detect modulations in speech resonances has been based on Teager's pioneering work on nonlinear modeling of human speech production [30]–[34].

In this paper, we provide an efficient solution to the fundamental problem of estimating the time-varying amplitude envelope and instantaneous frequency of a real-valued signal

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

that has both an AM and FM structure; we then apply these results to tracking modulations in speech resonances. Henceforth, we call $x(t)$ an AM–FM signal. It is a cosine of carrier frequency ω_c with a time-varying amplitude signal $a(t)$, an angle equal to the phase signal

$$\phi(t) = \omega_c t + \omega_m \int_0^t q(\tau)\,d\tau + \theta \tag{2}$$

and a time-varying instantaneous angular frequency signal

$$\omega_i(t) \triangleq \frac{d}{dt}\phi(t) = \omega_c + \omega_m q(t), \tag{3}$$

where $|q(t)| \leq 1$, ω_m is the maximum frequency deviation from ω_c, and $\theta = \phi(0)$ is some arbitrary constant phase offset. Note that two different information signals can be simultaneously transmitted in the amplitude $a(t)$ and the frequency $\omega_i(t)$ [or equivalently in the normalized frequency modulating signal $q(t)$]. Such AM–FM signals are very frequently used in communication systems. Next, we briefly outline our motivation for using them to model time-varying speech resonances.

By "speech resonances" we loosely refer to the oscillator systems formed by local cavities of the vocal tract emphasizing certain frequencies and de-emphasizing others during speech production. In linear speech modeling [2], [5], [6], [15], [16], [26], speech resonances, also

Manuscript received November 1991; revised July 1992. P. Maragos' research was supported in part by the National Science Foundation under Grant MIP-91-20624 and in part by a National Science Foundation Presidential Young Investigator Award under Grant MIP-86-58150 with matching funds from Bellcore, DEC, and Xerox. T. F. Quatieri's research was supported in part by the Department of the Air Force and in part by the Naval Submarine Medical Research Laboratory.

P. Maragos was with the Division of Applied Sciences, Harvard University, Cambridge, MA 02138, when this paper was first submitted. He is currently with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

J. F. Kaiser was with Bellcore, Morristown, NJ 07962, when this paper was first submitted; he is currently with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08855.

T. F. Quatieri is with the MIT Lincoln Laboratory, Lexington, MA 02173.

IEEE Log Number 9210950.

called "formants," are characterized by the poles of the transfer function of a linear filter modeling the vocal tract. Each pair of complex conjugate poles corresponds to a second-order resonator with an exponentially-damped cosine as impulse response

$$R_{\text{lin}}(t) = Ae^{-\sigma t} \cos(\omega_c t + \theta). \tag{4}$$

The formant frequency is ω_c, whereas $\sigma > 0$ controls the formant bandwidth. In the linear model, the model coefficients, and hence the formants, are assumed constant over each short-time analysis frame (about 10–30 msec, or 1–3 pitch periods). Thus, this classic approach assumes some local stationarity of the speech signal.

One of the implications of our work in this paper is that it now provides a possibility to extend this local stationarity assumption to a more refined model where variations of the frequency and amplitude of speech resonances can be modeled and detected at the smallest possible discretized time scale, i.e., the time scale of one sampling period. Experimental evidences motivating such an approach come from Teager's work [34] that provided indications and plausible explanations of how the speech resonances can change rapidly both in frequency and amplitude even within a single pitch period, possibly due to the rapidly-varying and separated speech airflow in the vocal tract. It is also known that slow time variations of the elements of simple second-order oscillators can result in amplitude or frequency modulation of the simple oscillator's cosine response [38]. Hence, since the effective air masses in vocal tract cavities and the effective cross-sectional areas of the airflow can vary rapidly, this could cause modulations of the air pressure and/or velocity field. In Section V we outline several mechanisms due to the physics of speech airflow that may explain such changes. All these considerations lead us to propose the following modulation model for each speech resonance:

$$\begin{aligned} R_{afm}(t) &= a(t) \cos[\phi(t)] \\ &= e^{-\sigma t} A(t) \cos\left(\omega_c t + \omega_m \int_0^t q(\tau)\, d\tau + \theta\right). \end{aligned} \tag{5}$$

Thus, the constant amplitude A and frequency ω_c in the linear resonator have now been replaced by a time-varying amplitude $A(t)$ and instantaneous frequency $\omega_i(t)$. This implies modeling the resonance with an exponentially-damped AM–FM model. The total speech signal $S(t)$ is then modeled as a sum of such AM–FM signals

$$S(t) = \sum_{k=1}^{K} a_k(t) \cos[\phi_k(t)] \tag{6}$$

where the subscript k refers to the kth resonance and K is the number of speech formants.

A fundamental problem is, given a speech signal over some time interval, to estimate the amplitude envelope $|a(t)|$ and the instantaneous frequency $\omega_i(t)$ of each resonance at each time t. Toward the solution of this problem, we first isolate individual resonances by bandpass filtering the speech signal around its formants. Then, we use efficient algorithms (presented in this paper) that can estimate the amplitude and frequency modulating signals of each resonance based on an "energy-tracking" operator. Specifically, Teager developed several tools for nonlinear speech processing such as the *energy operator*

$$\Psi_c[x(t)] \triangleq [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \tag{7}$$

for continuous-time signals $x(t)$ with $\dot{x} = dx/dt$ and its discrete-time counterpart

$$\Psi_d[x(n)] \triangleq x^2(n) - x(n-1)x(n+1) \tag{8}$$

for discrete-time signals $x(n)$, $n = 0, \pm 1, \pm 2, \cdots$. Both Ψ_c and Ψ_d are nonlinear and translation invariant. These operators were first introduced systematically by Kaiser [12], [13] and were shown to track the energy of simple harmonic oscillators. Namely, if $x(t) = A \cos(\omega_c t + \theta)$ is the displacement produced by a mass-spring undamped linear oscillator with mass m and spring constant k, then the total (kinetic plus potential) energy of this oscillator is equal to $(m\dot{x}^2 + kx^2)/2 = (m/2)A^2\omega_c^2$, i.e., proportional to the squared product of amplitude and frequency. Further, the operator Ψ_c applied to $x(t)$ can track the energy (per half-unit mass) of the source that produced the oscillation signal $x(t)$ because [13]

$$\Psi_c[A \cos(\omega_c t + \theta)] = (A\omega_c)^2 \tag{9}$$

for any constants A and ω_c.

The energy operators are also very useful for analyzing oscillation signals with time-varying amplitude and frequency. Specifically, in our earlier work [18], [19] we have shown that Ψ_c applied to the AM–FM signal of (1) can approximately estimate the squared product of the amplitude $a(t)$ and instantaneous frequency $\omega_i(t)$ signals; i.e.,

$$\Psi_c\left[a(t) \cos\left(\int_0^t \omega_i(\tau)\, d\tau + \theta\right)\right] \approx [a(t)\omega_i(t)]^2 \tag{10}$$

assuming that the signals $a(t)$ and $\omega_i(t)$ do not vary too fast (time rate of change of value) or too greatly (range of value) in time compared to the carrier frequency ω_c.

The first goal of this paper is to further separate the amplitude from the frequency signal in the output energy product of Ψ_c. Thus, we show in Section II how a combined use of the energy operator Ψ_c on the AM–FM signal and its derivative can lead to an elegant algorithm for separately estimating the amplitude $a(t)$ and frequency $\omega_i(t)$ signals. We call this an *energy separation algorithm* (ESA) because of the aforementioned dependence of the energy of an oscillator on the product of amplitude and frequency and because of the usage of energy-tracking operators.

Section III provides a similar ESA for discrete-time signals by using the discrete energy operator Ψ_d and approximating signal derivatives with 2-sample differences. Although the continuous and discrete-time cases have some

conceptual similarities, the mathematics are different and less intuitive in the discrete case; hence, the discrete case deserves its own exposition. In addition, because of the variety of possible discrete approximations to derivatives, there are several variations of discrete ESA's. For example, in Section IV we present a second discrete ESA that uses 3-sample differences in place of derivatives and comment on how its performance compares with the first discrete ESA.

Section V discusses the preliminary application of the discrete ESA to separate amplitude from frequency components in speech resonances, i.e., signals resulting from bandpass filtering speech vowels around their formants. While the major focus of the paper is the development of ESA's for arbitrary AM–FM signals, our main inspiration and motivation for this research has been based on the problem of tracking modulations in speech signals. Thus we precede the experimental results on speech by a brief discussion of theoretical and experimental evidences for the existence of such modulations in speech signals.

Finally, in Section VI we conclude by discussing various issues regarding the overall performance of ESA's for amplitude/frequency separation in AM–FM signals, which we have found impressive given the short-time nature of the nonlinear operators used and the relatively trivial complexity to implement them, as well as their implications for speech modeling.

II. Energy Separation for Continuous-Time Signals

In this section, we first present some closed-formula solutions for exact estimation of the constant amplitude and frequency of a cosine using the energy operator and then show that the same equations apply approximately to an AM–FM signal with time-varying amplitude and frequency. To simplify notation, we henceforth drop the subscripts c and d from the continuous and discrete energy operator symbols and use Ψ for both; the difference will always be clear from the context.

A. Constant Amplitude/Frequency Cosine

Consider a cosine $x(t) = A\cos(\omega_c t + \theta)$ with constant amplitude A and frequency $\omega_c > 0$ and its derivative

$$\dot{x}(t) = -A\omega_c \sin(\omega_c t + \theta). \tag{11}$$

Then (9) implies that

$$\Psi[\dot{x}(t)] = A^2\omega_c^4. \tag{12}$$

From (9) and (12) it follows that the constant frequency and the absolute amplitude of the cosine can be obtained from the following equations:

$$\omega_c = \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \tag{13}$$

$$|A| = \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}}. \tag{14}$$

At each time instant these equations use only the two instantaneous values of the output signals from the energy operator when the latter is applied to both the signal and its time derivative.

B. AM–FM Signals

Consider the real-valued AM–FM signal of (1)

$$x(t) = a(t)\cos[\phi(t)]. \tag{15}$$

We henceforth assume that $0 \le \omega_m < \omega_c$ and $|q(t)| \le 1$, which implies that for all t

$$0 < \omega_c - \omega_m \le \omega_i(t) < \omega_c + \omega_m < 2\omega_c. \tag{16}$$

In [19] we have shown that

$$\Psi[a\cos(\phi)] = (a\dot{\phi})^2 + \underbrace{0.5a^2\ddot{\phi}\sin(2\phi) + \Psi(a)\cos^2(\phi)}_{E(t)} \tag{17}$$

where

$$\dot{\phi} = \omega_i \quad \text{and} \quad \ddot{\phi} = \omega_m\dot{q}.$$

To obtain the result in (10), which generalizes (9) and extends it to time-varying amplitude and frequency signals, we henceforth view the term $E(t)$ as an *approximation error* and provide realistic conditions under which this error is negligible. Since most of the modulating signals with which we deal in this paper are narrowband and have an oscillatory nature, a meaningful (and tractable in the context of the energy operators) measure to quantify the error relative to the signal is the *relative maximum absolute error* defined as the ratio $E_{\max}/D_{\max}$, where $D = (a\dot{\phi})^2$ is the desired energy term in (17), and for an arbitrary signal $z(t)$ we denote

$$z_{\max} \triangleq \sup_t |z(t)|.$$

Thus, in the context of energy operators, we shall henceforth use an approximation $\approx$ to mean that

$$\Psi(x) = D + E \approx D \Leftrightarrow \frac{E_{\max}}{D_{\max}} \ll 1.$$

Hence, if the condition

$$\Psi(a)_{\max} + 0.5(a^2\ddot{\phi})_{\max} \ll (a\dot{\phi})^2_{\max} \tag{18}$$

holds, then (with a relative error $\ll 1$)

$$\Psi[a\cos(\phi)] \approx (a\dot{\phi})^2. \tag{19}$$

For amplitude/frequency separation we also need to apply Ψ to the AM–FM signal derivative $\dot{x} = y_1 - y_2$, where

$$\dot{x}(t) = \underbrace{\dot{a}(t)\cos[\phi(t)]}_{y_1(t)} - \underbrace{a(t)\dot{\phi}(t)\sin[\phi(t)]}_{y_2(t)}. \tag{20}$$

From (17) we obtain

$$\Psi[a\dot{\phi}\sin(\phi)] = a^2\dot{\phi}^4 - 0.5a^2\dot{\phi}^2\ddot{\phi}\sin(2\phi) + [a^2\Psi(\dot{\phi}) + \dot{\phi}^2\Psi(a)]\sin^2(\phi). \tag{21}$$

The desired term for energy separation is $a^2\dot{\phi}^4$, which can be combined with (19) to give simple equations for the amplitude and frequency signals. However, since

$$\Psi(y_1 - y_2) = \Psi(y_1) + \Psi(y_2) - 2\dot{y}_1\dot{y}_2 + y_1\ddot{y}_2 + \ddot{y}_1 y_2 \quad (22)$$

$\Psi(\dot{x})$ contains many other cross terms. To be able to analyze and compare the order of magnitude of these terms as well as find realistic assumptions under which the approximating condition (19) becomes valid, we next focus on the broad class of AM–FM signals with bandlimited amplitude/frequency modulating signals.

1) Bandlimited Amplitude/Frequency Modulating Signals: Let $z(t)$ be a real-valued signal with Fourier transform $Z(\omega)$. Consider its spectral absolute moments

$$\mu_{z,n} \triangleq \begin{cases} \frac{1}{\pi}\int_0^\infty \omega^n |Z(\omega)|\, d\omega, & \text{if } z(t) \text{ is aperiodic} \\ \omega_0^n \sum_{k=-\infty}^{\infty} |k|^n |\alpha_k|, & \text{if } z(t) = \sum_k \alpha_k e^{jk\omega_0 t} \end{cases} \quad (23)$$

for $n = 0, 1, 2, \cdots$ Note for the zeroth moment

$$\mu_z = \mu_{z,0}$$

that $z_{\max} = \mu_z$ if $z(t)$ is a cosine or its Fourier transform has a linear phase.

Consider now the following lemma for bandlimited signals whose proof can be found in [19].

Lemma 1 [19]: Let $z(t)$ be a real-valued continuous-time signal with Fourier transform $Z(\omega)$ whose spectral moments $\mu_{z,n}$ are finite for $n = 0, 1, 2$. Assume that $z(t)$ is bandlimited with highest frequency $\omega_z > 0$; i.e., $Z(\omega) = 0$ for $|\omega| > \omega_z$. Then

$$|z(t)| \le z_{\max} \le \mu_z \quad (24)$$

$$|\dot{z}(t)| \le \omega_z \mu_z \quad (25)$$

$$|\ddot{z}(t)| \le \omega_z^2 \mu_z \quad (26)$$

$$|\Psi[z(t)]| \le 2(\omega_z \mu_z)^2. \quad (27)$$

Let us now assume that $a(t)$ and $q(t)$ are bandlimited with highest frequencies ω_a and ω_f respectively, where ω_a, $\omega_f < \omega_c$. Then by Lemma 1 it follows that (see [19] for details)

$$|E(t)| \le (2\omega_a^2 + 0.5\omega_m\omega_f\mu_q)\mu_a^2. \quad (28)$$

To simplify the error formula, we henceforth assume that

$$\mu_a \approx a_{\max};\ \mu_q \approx q_{\max} = 1. \quad (29)$$

Then the main approximation (19) becomes valid (i.e., has negligible relative error) by assuming that

$$\frac{E_{\max}}{(a\omega_i)^2_{\max}} \le \frac{2\omega_a^2 + 0.5\omega_m\omega_f}{(\omega_c + \omega_m)^2} \ll 1. \quad (30)$$

In the special case of a *nonnegative* amplitude signal $a(t)$, we can assume without loss of generality (by ignoring constant scaling factors) that $a(t) = 1 + \kappa b(t)$ with $0 \le k \le 1$ and $|b(t)| \le 1$. Then [19]

$$|E(t)| \le \kappa\omega_a^2(\mu_b + 2\kappa\mu_b^2) + 0.5(1+\kappa)^2\omega_m\omega_f\mu_q \quad (31)$$

and the energy tracking in (19) has a negligible relative error if

$$\frac{E_{\max}}{(a\omega_i)^2_{\max}} \le \frac{\kappa\omega_a^2 + 0.5\beta\omega_f^2}{(\omega_c + \omega_m)^2} \ll 1 \quad (32)$$

where $\beta = \omega_m/\omega_f$ is the FM *modulation index*, κ is the AM index, and we assumed $\mu_b \approx 1$.

A simpler (and stronger) set of conditions that implies (30) or (32) is

$$\omega_a \ll \omega_c \text{ and } \kappa \ll 1$$

$$\omega_f \ll \omega_c \text{ and } \lambda = \frac{\omega_m}{\omega_c} \ll 1 \quad (33)$$

where $\lambda = \omega_m/\omega_c$ is the FM *modulation depth*. The above conditions formalize the intuitive idea that if the amplitude and the frequency signal do not vary too fast in time or too much compared to the carrier, then Ψ will track their squared product when applied to the AM–FM signal $x(t)$:

$$\Psi[x(t)] \approx a^2(t)\omega_i^2(t). \quad (34)$$

The error analysis is more complicated when we apply the energy operator to the derivative $\dot{x} = y_1 - y_2$ of the AM–FM signal, because $\Psi(\dot{x})$ contains many more terms than the desired term $a^2\omega_i^4$. One approach is to follow an approximate analysis where the dominant term is kept and the rest is ignored. Thus, notice from (20) and Lemma 1 that

$$(y_1)_{\max} \approx a_{\max}\omega_a \quad (35)$$

$$(y_2)_{\max} \approx a_{\max}(\omega_c + \omega_m). \quad (36)$$

Since $\omega_a \ll \omega_c$, the term y_2 has a much larger order of magnitude than y_1, where by "order of magnitude" we mean the physical order of magnitude of its maximum absolute value (see also the Appendix). Thus, by ignoring y_1, we can write

$$\dot{x} \approx -y_2 = -a\dot{\phi}\sin(\phi). \quad (37)$$

The AM–FM signal y_2 has an amplitude signal with bandwidth $\omega_a + \omega_f$. Hence, by replacing ω_a with $\omega_a + \omega_f$ in (30), (32) we arrive at the stricter conditions

$$2(\omega_a + \omega_f)^2 + 0.5\omega_m\omega_f \ll (\omega_c + \omega_m)^2$$
$$\kappa(\omega_a + \omega_f)^2 + 0.5\omega_m\omega_f \ll (\omega_c + \omega_m)^2, \quad (38)$$
$$\text{if } a(t) = 1 + \kappa b(t) \ge 0$$

which guarantee that

$$\Psi(y_2) \approx a^2\omega_i^4 \quad (39)$$

with a negligible relative error. Therefore,

$$\Psi[\dot{x}(t)] \approx a^2(t)\omega_i^4(t). \tag{40}$$

This result was derived by approximating the input signal $\dot{x} \approx -y_2$ and using the energy-tracking capability of Ψ as in (19). An alternative approach is not to approximate the input $\dot{x}$, but instead to find the order of magnitude of all the error terms in the output signal $\Psi(\dot{x})$ and show that they are much smaller than the order of magnitude of the desired term $a^2\omega_i^4$. In the Appendix we provide a detailed analysis for this second approach.

By combining now (34) and (40) we obtain

$$\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx \omega_i(t) \tag{41}$$

$$\frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)|. \tag{42}$$

We call (41), (42) the *continuous energy separation algorithm* (CESA). Thus, the CESA provides estimates of the time-varying instantaneous frequency signal $\omega_i(t) \geq 0$ and of the *amplitude envelope* $|a(t)|$ of an AM–FM signal, given the assumptions (33). Note that if $x(t) = A\cos(\omega_c t + \theta)$ is a cosine with constant amplitude/frequency, then the CESA yields the exact solution $\omega_i(t) = \omega_c$ and $|a(t)| = |A|$.

For the validity of the approximate results (10), (40), and the CESA, it is assumed that we deal only with signals x for which $\Psi(x)$ and $\Psi(\dot{x})$ are *nonnegative* signals. From (17) it follows that a *sufficient* condition for $\Psi[x(t)] \geq 0$ for all t is, assuming that $a(t) = 1 + \kappa b(t) > 0$ with $\kappa < 1$,

$$\Psi(a)_{\max} + 0.5\omega_m(1+\kappa)^2\dot{q}_{\max} \leq (1-\kappa)^2(\omega_c - \omega_m)^2. \tag{43}$$

As we have shown in [19], there are broad classes of AM–FM signals x with (often extremely) large amounts of AM or FM modulation (i.e., large κ and λ) for which $\Psi(x) \geq 0$. If $\Psi(a) > 0$, then another sufficient condition [simpler than (43)] for nonnegativity of $\Psi[a\cos(\phi)]$ is $\omega_m\dot{q}_{\max} \leq 2(\omega_c - \omega_m)^2$. There is a large class of amplitude modulating signals a such that $\Psi(a) \geq 0$. For example, this class includes i) cosines of constant frequency, ii) exponentials $e^{\sigma t}$ since $\Psi(e^{\sigma t}) = 0$, and iii) linear trends $st + c$ since $\Psi(st + c) = s^2$, as well as all finite products of signals from any of these three classes.

Note also that at times t_0 of zero energy, i.e., when $\Psi[x(t_0)] = 0$, we approximately have zero amplitude $a(t_0) = 0$ if the frequency $\omega_i(t)$ is assumed to always be positive. At such time instants, however rare they may be, we need additional information to estimate $\omega_i(t_0)$. Given the assumption that $\omega_i(t)$ varies more slowly than $x(t)$, one approach is to interpolate the missing $\omega_i(t_0)$ value from its surrounding neighbors.

2) Special Cases of Amplitude/Frequency Modulating Signals: The CESA was derived by assuming that the amplitude and frequency information-carrying signals a and q are bandlimited. However, there are also other special cases of AM–FM signals, where bandlimitedness is not a true or an appropriate (for Lemma 1) assumption. Then, it may be possible to show that the CESA will still yield approximately correct solutions provided that the information signals do not vary too fast or too much with time compared to the carrier. We next analyze two such cases.

Cosine with Exponential Amplitude: For any real σ, let

$$x(t) = Ae^{\sigma t}\cos(\omega_c t + \theta). \tag{44}$$

This can be viewed as a special case of an AM–FM signal with time-varying amplitude $a(t) = Ae^{\sigma t}$ and constant frequency. Now

$$\dot{x}(t) = Ae^{\sigma t}[-\omega_c \sin(\omega_c t + \theta) + \sigma\cos(\omega_c t + \theta)]. \tag{45}$$

By applying Ψ to x and $\dot{x}$ we have

$$\Psi[x(t)] = A^2\omega_c^2 e^{2\sigma t} \tag{46}$$

$$\Psi[\dot{x}(t)] = A^2\omega_c^4 e^{2\sigma t}\left(1 + \frac{\sigma^2}{\omega_c^2}\right). \tag{47}$$

Then assuming that $(\sigma/\omega_c)^2 \ll 1$ yields

$$\sqrt{\frac{\Psi(\dot{x})}{\Psi(x)}} = \omega_c\sqrt{1 + \frac{\sigma^2}{\omega_c^2}} \approx \omega_c \tag{48}$$

$$\frac{\Psi(x)}{\sqrt{\Psi(\dot{x})}} = \frac{|A|e^{\sigma t}}{\sqrt{1 + \dfrac{\sigma^2}{\omega_c^2}}} \approx |A|e^{\sigma t}. \tag{49}$$

Thus, if the exponential rate $|\sigma|$ is much smaller than ω_c, we can use the CESA to approximately estimate the amplitude and frequency. The relative approximation error is about $\sigma^2/2\omega_c^2$. Note that if $\sigma < 0$, then $|\sigma|$ is the 3 dB bandwidth of the exponential $e^{\sigma t}$. Hence, the condition $(\sigma/\omega_c)^2 \ll 1$ is of the same spirit as the condition $(\omega_a/\omega_c)^2 \ll 1$ for truly bandlimited amplitude signals $a(t)$.

FM/Linear (Chirp): Consider the following FM signal over a finite-time interval

$$x(t) = A\cos\underbrace{\left(\omega_c t + \omega_m\left(\frac{t^2}{L} - t\right) + \theta\right)}_{\phi(t)}, \qquad t \in [0, L] \tag{50}$$

with linear instantaneous frequency $\omega_i(t) = \omega_c + \omega_m(2t/L - 1)$. Then

$$\dot{x}(t) = -A\omega_i(t)\sin\left(\omega_c t + \omega_m\left(\frac{t^2}{L} - t\right) + \theta\right). \tag{51}$$

From (17) it follows that

$$\Psi(x) = A^2[\omega_i^2 + \omega_m \sin(2\phi)/L] \quad (52)$$

$$\approx A^2\omega_i^2(t) \quad (53)$$

$$\Psi(\dot{x}) = A^2\left[\omega_i^4 - \omega_i^2\frac{\omega_m}{L}\sin(2\phi) + \frac{4\omega_m^2}{L^2}\sin^2(\phi)\right] \quad (54)$$

$$\approx A^2\omega_i^4(t) \quad (55)$$

where both approximations assumed that

$$4\frac{\omega_m}{L} \ll (\omega_c + \omega_m)^2. \quad (56)$$

Under this condition, the CESA can be used to approximately estimate the amplitude and frequency of the FM/Linear signal.

III. Discrete-Time Energy Separation Algorithm-1 (DESA-1)

In this section we derive an algorithm for estimating the amplitude envelope and the instantaneous frequency of a discrete-time AM–FM signal by using the discrete-time energy operator. The basic inspiration for a discrete algorithm comes again from the case of a cosine with constant amplitude and frequency. Thus we first discuss this constant case and then we focus on the general AM–FM case.

A. Cosine with Constant Amplitude/Frequency

If $x(n)$ is a sampled version of a continuous-time signal and we replace derivatives $\dot{x}$ with 2-sample backward (or forward) differences $[x(n) - x(n-1)]/T$ where T is the sampling period, then [19] the continuous-time energy operator reduces (up to one sample shift) to the following discrete version

$$\Psi[x(n)] = [x^2(n) - x(n+1)x(n-1)]/T^2.$$

Consider a constant amplitude/frequency discrete-time cosine

$$x(n) = A\cos(\Omega_c n + \theta) \quad (57)$$

where $\Omega_c = \omega_c T$ and $\omega_c < \pi/T$. Then [12]

$$\Psi[x(n)] = A^2\sin^2(\Omega_c)/T^2 \quad (58)$$

which can be rewritten as

$$\Psi[x(n)] = A^2\omega_c^2\left(\frac{\sin\Omega_c}{\Omega_c}\right)^2. \quad (59)$$

In this form, a comparison with the continuous case (9) is direct; the discrete case has the additional $(\sin\Omega_c/\Omega_c)^2$ factor which attenuates the result by a known and compensatable amount. Thus, the two results are similar. Since the variable Ω_c carries the information about T, we can assume $T = 1$ (thus simplifying many of the expressions and formulas) and then insert the proper sampling period or frequency at the end to scale the answer according to the relation $f_c = \Omega_c/(2\pi T)$, where f_c is the frequency in Hz. In the remainder of this paper we have therefore assumed that $T = 1$ in the expression of Ψ and hence

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1).$$

Note that $T = 1$ is also the correct assumption for signals that are inherently defined only for discrete time.

Now consider the backward difference approximation for the first derivative of $x(t)$:

$$y(n) = x(n) - x(n-1)$$
$$= A[\cos(\Omega_c n + \theta) - \cos(\Omega_c(n-1) + \theta)]$$
$$= -2A\sin(\Omega_c/2)\sin(\Omega_c n + \theta - \Omega_c/2) \quad (60)$$

given that

$$\cos(\alpha) - \cos(\beta) = 2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\beta-\alpha}{2}\right).$$

Now

$$\Psi[y(n)] = 4A^2\sin^2(\Omega_c/2)\sin^2(\Omega_c). \quad (61)$$

Note that

$$\frac{\Psi[y(n)]}{2\Psi[x(n)]} = 2\sin^2(\Omega_c/2) = 1 - \cos(\Omega_c). \quad (62)$$

Hence, the constant frequency and absolute amplitude can be found from the following equations:

$$\Omega_c = \arccos\left(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\right) \quad (63)$$

$$|A| = \sqrt{\frac{\Psi[x(n)]}{\sin^2(\Omega_c)}} = \sqrt{\frac{\Psi[x(n)]}{1 - \cos^2(\Omega_c)}} \quad (64)$$

$$= \sqrt{\frac{\Psi[x(n)]}{1 - \left(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\right)^2}} \quad (65)$$

Note that we can also find the frequency and amplitude by using the above equations and replacing $y(n)$ with the forward difference $x(n+1) - x(n)$.

B. AM–FM Signals

Consider a discrete-time real-valued AM–FM signal

$$x(n) = a(n)\cos\Big[\underbrace{\Omega_c n + \Omega_m\int_0^n q(m)\,dm + \theta}_{\phi(n)}\Big]. \quad (66)$$

We define its instantaneous frequency by

$$\Omega_i(n) \triangleq \frac{d}{dn}\phi(n) = \Omega_c + \Omega_m q(n). \quad (67)$$

Both the differentiation d/dn in (67) and the integration $\int dm$ in (66) treat the integer time indexes n and m sym-

bolically as continuous variables. Thus, in the discrete FM case, $q(\cdot)$ is assumed to be a known mathematical function with a computable integral. Note that the continuous-time angular frequencies ω_c, ω_m, and ω_i (in radians/s) have been replaced by their discrete-time counterparts Ω_c, Ω_m, and Ω_i (in radians/sample). All frequency variables in this paper are assumed nonnegative, and all discrete-time frequencies are assumed to be $< \pi$. In addition, we assume that $|q(n)| \leq 1$ and $\Omega_m \leq \Omega_c$, which implies that for all n

$$0 < \Omega_c - \Omega_m \leq \Omega_i(n) \leq \Omega_c + \Omega_m < \pi. \tag{68}$$

Now for any $a(n)$ and $\phi(n)$

$$\Psi[x(n)] = a^2(n)\Psi[\cos(\phi(n))] + \Psi(a(n))[\cos^2(\phi(n)) - \Psi(\cos\phi(n))]. \tag{69}$$

If we assume (as shown later for slowly-varying $\Omega_i(n)$ compared to Ω_c) that

$$\Psi[\cos(\phi(n))] \approx \sin^2[\Omega_i(n)] \tag{70}$$

then

$$\Psi[x(n)] \approx a^2(n)\sin^2[\Omega_i(n)] + \Psi(a(n))[\cos^2(\phi(n)) - \sin^2(\Omega_i(n))] \tag{71}$$

$$\approx a^2(n)\sin^2[\Omega_i(n)], \tag{72}$$

$$\text{if } \Psi(a)_{\max} \ll [a_{\max}\sin(\Omega_i)_{\max}]^2$$

where in general

$$[\sin(\Omega_i)]_{\max} = \begin{cases} \sin(\Omega_c + \Omega_m) & \text{if } \Omega_c + \Omega_m \leq \pi/2 \\ 1 & \text{if } |\Omega_c - \pi/2| \leq \Omega_m \\ \sin(\Omega_c - \Omega_m) & \text{if } \Omega_c - \Omega_m > \pi/2 \end{cases} \tag{73}$$

To find appropriate classes of signals for which the above approximations become valid and tractable, we next focus on the case of bandlimited modulating signals.

1) Bandlimited Amplitude/Frequency Modulating Signals: Given a real-valued signal $z(n)$ with Fourier transform $Z(\Omega)$ define its mean absolute spectral value by

$$M_z \triangleq \begin{cases} \frac{1}{\pi}\int_0^{\pi} |Z(\Omega)|\, d\Omega, & \text{if } z \text{ is aperiodic} \\ \sum_{k=0}^{N-1} |\alpha_k|, & \text{if } z(n) = \sum_{k=0}^{N-1} \alpha_k e^{j2\pi kn/N}. \end{cases} \tag{74}$$

Note that $z_{\max} = M_z$ if $z(n)$ is a cosine or has a linear Fourier phase.

To find bounds on the output of the energy operator processing bandlimited signals we first need the following lemma.

Lemma 2 [19]: Let $z(n)$ be a real-valued discrete-time signal with a finite mean absolute spectral value M_z. Assume that $z(n)$ is bandlimited with highest frequency $\Omega_z \geq 0$, i.e., $Z(\Omega) = 0$ for $\Omega_z < |\Omega| \leq \pi$. Then

$$|z(n)| \leq z_{\max} \leq M_z \tag{75}$$

$$|z(n) - z(n-1)| \leq 2\sin(\Omega_z/2)M_z \tag{76}$$

$$|\Psi(z(n))| \leq 8\sin^2(\Omega_z/2)M_z^2 \tag{77}$$

Now for the AM–FM signal assume that the amplitude signal $a(n)$ is bandlimited with bandwidth $\Omega_a < \Omega_c$ and that $a_{\max} \approx M_a$.

For the frequency $\Omega_i(n)$ we assume that it is generally a nonnegative signal that can be expressed as a finite linear combination of cosines. This class of frequency signals is quite broad since it includes all real-valued periodic signals, which (via the DFT) can always be expressed as a finite sum of cosines. Therefore, with no loss of generality,[1] we can write

$$\Omega_i(n) = \Omega_c + \sum_{k=1}^{K} \Omega_{m,k}\cos(\Omega_{f,k}n + \theta_k) \tag{78}$$

with $\Omega_c > 0$ and all $\Omega_{m,k} > 0$. If $\Omega_i(n)$ is periodic with period $N+1$, then $K \leq \lfloor (N+1)/2 \rfloor$, and the dc component of the instantaneous frequency signal is

$$\Omega_c = \frac{1}{N+1}\sum_{n=0}^{N} \Omega_i(n). \tag{79}$$

We can also express $\Omega_i(n)$ in the standard form (67) by writing

$$\Omega_i(n) = \Omega_c + \Omega_m q(n) \tag{80}$$

where the maximum frequency deviation is

$$\Omega_m = \sum_{k=1}^{K} \Omega_{m,k} \tag{81}$$

and hence $|q(n)| \leq 1$. The phase signal corresponding to (78) is

$$\phi(n) = \Omega_c n + \sum_{k=1}^{K} \frac{\Omega_{m,k}}{\Omega_{f,k}}\sin(\Omega_{f,k}n + \theta_k) + \theta. \tag{82}$$

The bandwidth of $\Omega_i(n)$ is equal to

$$\Omega_f = \max_{k=1}^{K} \Omega_{f,k}. \tag{83}$$

[1] Assuming that $\Omega_i(n)$ is periodic presents no loss of generality in applying the energy operator/separation algorithms on a short-time basis to speech and many other discrete-time signal classes. Namely, any short-time segment of a discrete-time signal or any finite-extent signal can be repeated and viewed as periodic. Any real-valued periodic discrete-time signal can be expressed as a finite linear combination of cosines; adjusting their phase offsets yields $\Omega_{m,k} > 0$, and if the signal is nonnegative its dc component is $\Omega_c > 0$.

Due to the specific nature of $\phi(n)$ the following general approximations result:

$$\phi(k) + \phi(m) \approx 2\phi\left(\frac{k+m}{2}\right), \quad \text{if } \Omega_f |k-m| \ll 2 \tag{84}$$

$$\phi(k) - \phi(m) \approx (k-m)\Omega_i\left(\frac{k+m}{2}\right), \quad \text{if } \Omega_f |k-m| \ll 2. \tag{85}$$

Henceforth, we assume that

$$\Omega_f \ll 1 \tag{86}$$

which means that the frequency signal $\Omega_i(n)$ has a small bandwidth. Then by setting $m = n + 1$ and $k = n - 1$ in (84), (85) we obtain

$$\begin{aligned}\cos[\phi(n+1)]\cos[\phi(n-1)] &= \tfrac{1}{2}(\cos[\phi(n+1)+\phi(n-1)] + \cos[\phi(n+1) - \phi(n-1)]) \\ &\approx \underbrace{\tfrac{1}{2}(\cos[2\phi(n)] + \cos[2\Omega_i(n)])}_{\cos^2[\phi(n)] - \sin^2[\Omega_i(n)]}.\end{aligned} \tag{87}$$

Therefore,

$$\Psi[\phi(n)] \approx \sin^2[\Omega_i(n)]. \tag{88}$$

Hence, the approximation (70) becomes valid due to assumption (86).

From Lemma 2 it follows [19] that, the approximation in (72) also becomes valid if we further assume for the bandlimited $a(n)$ that

$$\begin{aligned}8\sin^2(\Omega_a/2) &\ll [\sin^2(\Omega_i)]_{\max} \\ 4\kappa\sin^2(\Omega_a/2) &\ll [\sin^2(\Omega_i)]_{\max},\end{aligned} \quad \text{if } a(n) = 1 + \kappa b(n) \geq 0. \tag{89}$$

Now, consider the backward difference

$$\begin{aligned}y(n) &= x(n) - x(n-1) \\ &= \underbrace{a(n)c(n)}_{D(n)} + \underbrace{[a(n) - a(n-1)]\cos[\phi(n-1)]}_{E(n)}\end{aligned} \tag{90}$$

where

$$c(n) = \cos[\phi(n)] - \cos[\phi(n-1)] \tag{91}$$

$$= 2\sin\left[\frac{\phi(n) + \phi(n-1)}{2}\right] \cdot \sin\left[\frac{\phi(n-1) - \phi(n)}{2}\right]. \tag{92}$$

From (84)–(86) we obtain

$$c(n) \approx -2\sin[\Omega_i(n-0.5)/2]\sin[\phi(n-0.5)]. \tag{93}$$

Lemma 2 implies that the order of magnitude of the terms in (90) are

$$\begin{aligned}D_{\max} &\approx 2\sin(\Omega_i/2)_{\max} a_{\max} \\ E_{\max} &\approx 2\sin(\Omega_a/2) a_{\max}\end{aligned} \tag{94}$$

Due to assumption (89), D is the dominant term because its order of magnitude is much larger than that of E. Hence, ignoring the term E,

$$y(n) \approx -2a(n)\sin[\Omega_i(n-0.5)/2]\sin[\phi(n-0.5)]. \tag{95}$$

Assuming that

$$\Omega_m \ll \Omega_c \tag{96}$$

and using the standard series expansions for sin( ) and cos( ) together with the first-order approximation $(1 + v)^p \approx 1 + pv$ if $|v| \ll 1$ yields, for any real r,

$$\sin[r\Omega_i(n)] \approx \sin(r\Omega_c) + r\Omega_m \cos(r\Omega_c) q(n). \tag{97}$$

Thus, the approximate amplitude signal of $y(n)$ is essentially bandlimited with highest frequency $\Omega_a + \Omega_f$. Hence, if we make the assumption [stronger than (89)]

$$\begin{aligned}8\sin^2[(\Omega_a + \Omega_f)/2] &\ll [\sin^2(\Omega_i)]_{\max} \\ 4\kappa\sin^2[(\Omega_a + \Omega_f)/2] &\ll [\sin^2(\Omega_i)]_{\max}\end{aligned} \quad \text{if } a(n) = 1 + \kappa b(n) \geq 0 \tag{98}$$

then

$$\Psi[y(n)] \approx 4a^2(n)\sin^2[\Omega_i(n-0.5)/2] \cdot \sin^2[\Omega_i(n-0.5)]. \tag{99}$$

To proceed further, one approach is to ignore the half-sample shift, since

$$|\Omega_i(n-0.5) - \Omega_i(n)| \leq \frac{\Omega_m\Omega_f}{2} \ll \Omega_c. \tag{100}$$

Then, we can set $\Omega_i(n) \approx \Omega_i(n-0.5)$ and combine (72) and (99) to derive the following *discrete energy separation algorithm* (DESA) for estimating the instantaneous frequency and envelope:

$$\arccos\left(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\right) \approx \Omega_i(n) \tag{101}$$

$$\sqrt{\frac{\Psi[x(n)]}{1 - \left(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\right)^2}} \approx |a(n)|. \tag{102}$$

We call this algorithm DESA-1a, where "1" implies the approximation of derivatives with a single sample differ-

ence and "a" refers to the usage of asymmetric difference. We have experimentally tested DESA-1a and found that it generally works quite well in estimating the amplitude and frequency signals. (Numerical results are given later). However, since we used an asymmetric difference for approximating the derivative $\dot{x}$, a further improvement may result (as discussed later) if we symmetrize the term $\Psi[x(n) - x(n-1)]$ by averaging it with $\Psi[x(n+1) - x(n)]$. Thus, if we repeat the analysis done for the backward difference y for the forward difference z, we obtain

$$z(n) = x(n+1) - x(n) = y(n+1) \tag{103}$$

$$\approx 2a(n) \sin\left[\frac{\Omega_i(n+\frac{1}{2})}{2}\right] \sin[\phi(n+0.5)] \tag{104}$$

and applying the energy operator

$$\Psi[z(n)] \approx 4a^2(n) \sin^2[\Omega_i(n+0.5)/2] \cdot \sin^2[\Omega_i(n+0.5)]. \tag{105}$$

By averaging the results in (99) and (105) and assuming that the shifts by $+\frac{1}{2}$ and $-\frac{1}{2}$ sample approximately cancel out, we have

$$\frac{\Psi[y(n)] + \Psi[z(n)]}{2} \approx 4a^2(n) \sin^2[\Omega_i(n)/2] \sin^2[\Omega_i(n)]. \tag{106}$$

Thus, the action of Ψ on asymmetric derivatives is partially "symmetrized" by averaging the action of Ψ on two opposite asymmetric derivatives. Then combining (72) and (106) yields another DESA given by the formulas

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

$$\arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right) \approx \Omega_i(n) \tag{107}$$

$$\sqrt{\frac{\Psi[x(n)]}{1 - \left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2}} \approx |a(n)| \tag{108}$$

We call this the DESA-1 algorithm. The frequency estimation part works as long as $0 < \Omega_i(n) < \pi$, since the principal value of the arccos (v) function assumes that $v \in [0, \pi]$. Thus, the DESA-1 algorithm can estimate instantaneous frequencies up to $\frac{1}{2}$ the sampling frequency. Note that we can replace the left side of (107) with its equivalent expression

$$2 \arcsin\left(\sqrt{\frac{\Psi[y(n)] + \Psi[y(n+1)]}{8\Psi[x(n)]}}\right) \approx \Omega_i(n) \tag{109}$$

However, the formula using arcsin( ) is computationally more complex than the formula using arccos( ) because it requires an additional square-root operation per sample.

For the validity of the approximate results (72), (99), and (105), and the DESA-1, it is assumed in this paper that we deal with signals $x(n)$ such that $\Psi[x(n)] \geq 0$ and $\Psi[x(n) - x(n-1)] \geq 0$ for all n. As discussed for the continuous-time case, this is satisfied by many classes of AM-FM signals [19]. For the present work it suffices to say that, in most of our experiments with noiseless AM-FM and bandpass filtered speech signals, we have very rarely encountered a negative $\Psi[x(n)]$, and in most such cases the negative value appeared to be due to round-off errors. Note also that, by (72), if for some n_0 we have $\Psi[x(n_0)] = 0$, then assuming that $\Omega_i(n_0) \neq 0$, we have $a(n_0) \approx 0$; also, $\Psi[y(n_0)] \approx 0$. Thus, at such time instants we cannot estimate the frequency $\Omega_i(n)$ using the DESA-1. In our implementations, whenever we encountered such a situation we set the amplitude equal to zero and estimated the frequency from its value at the previous sample, i.e., set $\Omega_i(n_0) = \Omega_i(n_0 - 1)$. This is in accordance with our general assumption that Ω_i varies more slowly than the signal x. An alternative implementation would be to interpolate the frequency signal value at such time instants from its neighbors.

Fig. 1 shows that the application of DESA-1 to an AM-FM signal, whose amplitude and frequency modulating signals were single cosines, results in a successful approximate estimation of the amplitude and frequency signals with relatively small errors. Note that both the AM and the FM parts of this AM-FM signal have a large amount of modulation, i.e., 50% AM since $\kappa = 0.5$ and 20% FM since $\lambda = \Omega_m/\Omega_c = 0.2$. Despite these large amounts of modulation, the DESA performs well. Obviously, its performance is even better at lower amounts of modulation. Also note that Fig. 1(b) shows the square root of the output of the energy operator when applied to the original AM-FM signal; this output signal is approximately the product of the amplitude and the sine of the frequency signal. Then the DESA-1 separates these two information signals. Finally, we have found that DESA-1 performs very similarly to DESA-1a, although the latter almost always yields somewhat larger estimation errors. Numerical comparisons between the errors of DESA-1 and DESA-1a are given in Section IV.

Concluding, we note that it is very simple to implement DESA-1 since it only requires a few simple operations per output sample and involves a very short window of samples around the time instant at which we estimate the amplitude and frequency. Details on the computational complexity of DESA-1 are discussed later in Section VI-A.

2) Bandlimited-Amplitude/Linear-Frequency Modulating Signals:

Consider now a discrete AM-FM/Linear signal $x(n) = a(n) \cos[\phi(n)]$ over a finite time interval, i.e., a chirp signal with a time-varying amplitude $a(n)$ and linearly-varying instantaneous frequency

$$\Omega_i(n) = \Omega_c + \Omega_m\left(\frac{2n}{N} - 1\right), \quad n = 0, 1, \cdots, N. \tag{110}$$

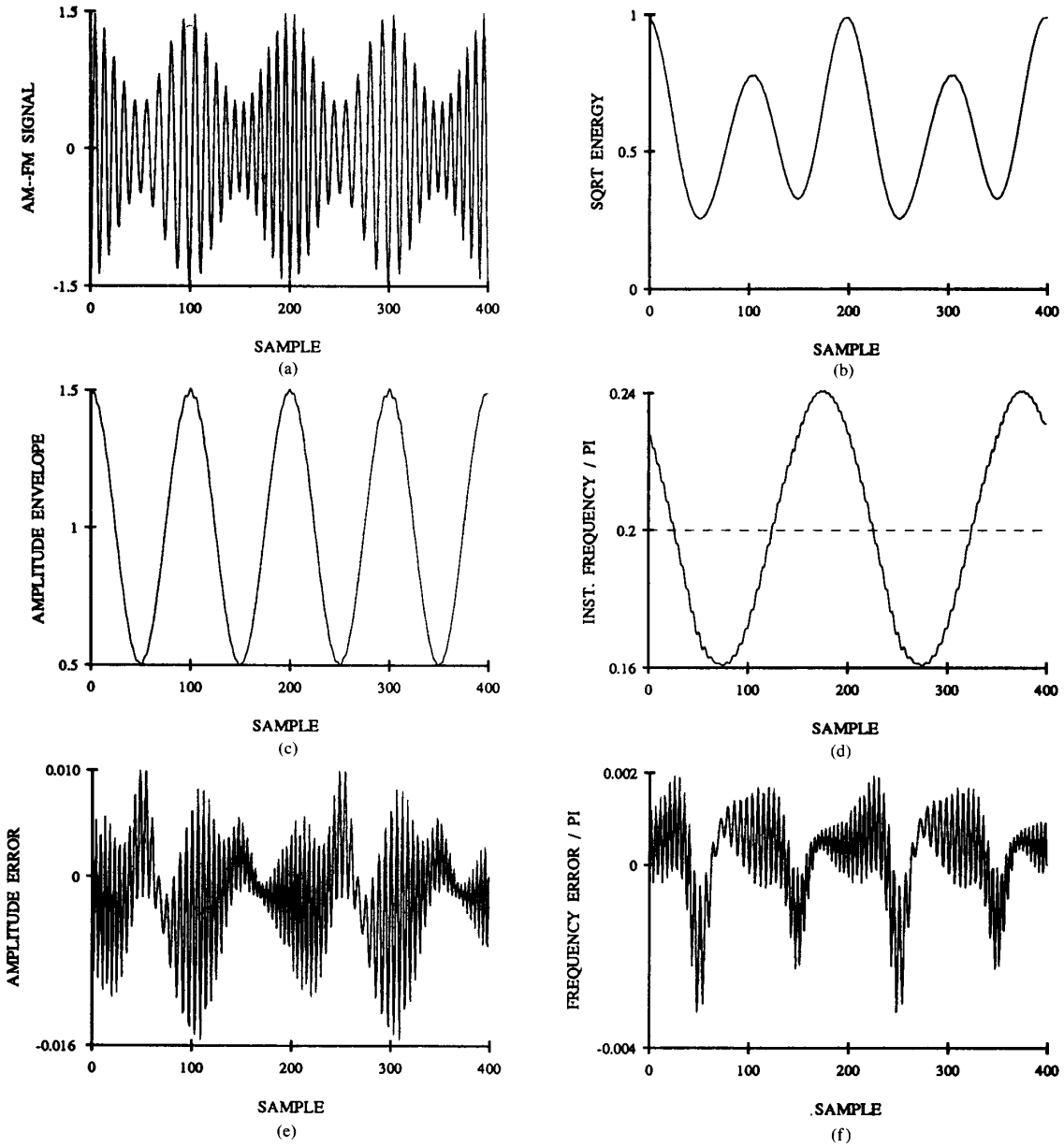


Fig. 1. (a) AM-FM/Cosine signal $x(n) = [1 + 0.5 \cos(\pi n/50)] \cos[\pi n/5 + 4 \sin(\pi n/100 + \pi/4)]$. (b) $\sqrt{\Psi}[x(n)]$. (c) Estimated amplitude envelope using DESA-1. (d) Estimated instantaneous frequency (shown as fraction of π); the dotted line shows Ω_c/π. (e) Error in amplitude estimation. (f) Error in frequency estimation.

Its quadratic phase signal is

$$\phi(n) = \Omega_c n + \Omega_m \left(\frac{n^2}{N} - n\right) + \phi(0). \quad (111)$$

By repeating $\Omega_i(n)$ every $N + 1$ samples one can view it as a periodic signal. However, although we can represent this linear $\Omega_i(n)$ over the interval $[0, N]$ as a combination of cosines (via DFT), it is not effectively bandlimited in the sense of not having a highest frequency $\ll 1$. Thus, we use a different approach than in the case of bandlimited frequency modulating signals.

First, note that for any integers k, m

$$\phi(k) + \phi(m) = 2\phi\left(\frac{k+m}{2}\right) + \frac{\Omega_m}{2N}(k-m)^2 \quad (112)$$

$$\phi(k) - \phi(m) = (k-m)\Omega_i\left(\frac{k+m}{2}\right). \quad (113)$$

By setting $k = n + 1$ and $m = n - 1$ it follows that

$$\begin{aligned}\cos[\phi(n-1)]\cos[\phi(n+1)] &= \frac{1}{2}(\cos[\phi(n+1)+\phi(n-1)] \\ &\quad + \cos[\phi(n+1)-\phi(n-1)]) \\ &= \frac{1}{2}\left(\cos\left[2\phi(n)+\frac{2\Omega_m}{N}\right]+\cos[2\Omega_i(n)]\right) \\ &= \cos^2[\phi(n)] - \sin^2[\Omega_i(n)] - \sin\left[2\phi(n)+\frac{\Omega_m}{N}\right] \\ &\quad \cdot \sin\left(\frac{\Omega_m}{N}\right). \end{aligned} \tag{114}$$

Hence, if we assume

$$\sin\left(\frac{\Omega_m}{N}\right) \ll \sin^2(\Omega_i)_{\max} \tag{115}$$

we obtain the approximation $\Psi[\phi(n)] \approx \sin^2[\Omega_i(n)]$. Note that (115) implies

$$\frac{\Omega_m}{N} \ll 1. \tag{116}$$

Then, by further assuming that $a(n)$ is bandlimited with bandwidth Ω_a satisfying (89) we obtain

$$\Psi[a(n)\cos(\phi(n))] \approx a^2(n)\sin^2[\Omega_i(n)]. \tag{117}$$

Now consider the backward difference

$$\begin{aligned} y(n) &= x(n) - x(n-1) \\ &= a(n)c(n) + [a(n) - a(n-1)]\cos[\phi(n-1)] \end{aligned} \tag{118}$$

where

$$c(n) = \cos[\phi(n)] - \cos[\phi(n-1)] \tag{119}$$

$$= -2\sin\left[\frac{\Omega_i(n-0.5)}{2}\right]\sin\left[\phi(n-0.5)+\frac{\Omega_m}{4N}\right]. \tag{120}$$

By (89), the term $a(n)c(n)$ has a larger order of magnitude than the residual $y(n) - a(n)c(n)$. Hence, we can set $y(n) \approx a(n)c(n)$. Assuming further that

$$\Omega_m \ll \Omega_c \tag{121}$$

and using the approximation (97) leads us to conclude that the effective bandwidth of $a(n)\sin(\Omega_i(n-0.5)/2)$ is equal to Ω_a. Therefore, from (89), (115), and (117) it follows that

$$\Psi[y(n)] \approx 4a^2(n)\sin^2[\Omega_i(n-0.5)/2] \cdot \sin^2[\Omega_i(n-0.5)]. \tag{122}$$

Note that

$$\Omega_i(n) - \Omega_i(n-0.5) = \frac{\Omega_m}{2N} \ll \Omega_c. \tag{123}$$

Hence, we can set $\Omega_i(n) \approx \Omega_i(n-0.5)$ and apply the asymmetric algorithm DESA-1a of (101), and (102) to find both $\Omega_i(n)$ and $|a(n)|$. In addition, our extensive numerical comparisons (discussed later) have shown that a further improvement almost always results if we apply the symmetrized algorithm DESA-1.

Fig. 2 shows the good amplitude/frequency separation and tracking capabilities of the DESA-1 in the case of an AM-FM signal with a sinusoidally-varying amplitude and linearly-varying frequency components. The percent of both AM and FM is 25% in this example. Despite the discontinuity in the slope of the frequency signal, we see that DESA-1 performs quite well in separating and tracking the amplitude and frequency signals.

IV. Discrete-Time Energy Separation Algorithm-2 (DESA-2)

In this section, we develop an alternative DESA that avoids the previous half-sample shifts in the estimated frequency signal by using a symmetric difference to approximate the first derivative $\dot{x}(t)$. We first start from the simple case of a cosine with constant amplitude/frequency that inspires the specific algorithm. Then we show that the algorithm also works for general AM-FM signals.

A. Cosine with Constant Amplitude/Frequency

Consider the cosine

$$x(n) = A\cos(\Omega_c n + \theta) \tag{124}$$

and the 3-sample *symmetric* difference

$$\begin{aligned} s(n) &= [(x(n+1) - x(n)) + (x(n) - x(n-1))]/2 \\ &= [x(n+1) - x(n-1)]/2 \\ &= A[\cos(\Omega_c(n+1)+\theta) \\ &\quad - \cos(\Omega_c(n-1)+\theta)]/2 \\ &= -A\sin(\Omega_c)\sin(\Omega_c n + \theta) \end{aligned} \tag{125}$$

$s(n)$ is the simplest symmetric form for the approximation of the first derivative $\dot{x}$ that gives the estimate of the derivative at a sample point. Then

$$\Psi[s(n)] = A^2\sin^4(\Omega_c). \tag{126}$$

By combining (126) with (58) we obtain

$$\sin^2(\Omega_c) = \frac{1-\cos(2\Omega_c)}{2} = \frac{\Psi[s(n)]}{\Psi[x(n)]} \tag{127}$$

$$A^2 = \frac{\Psi^2[x(n)]}{\Psi[s(n)]}. \tag{128}$$

Hence, the constant frequency and absolute amplitude can be obtained from the following equations:

$$\Omega_c = \arcsin\left(\sqrt{\frac{\Psi[x(n+1) - x(n-1)]}{4\Psi[x(n)]}}\right) \tag{129}$$

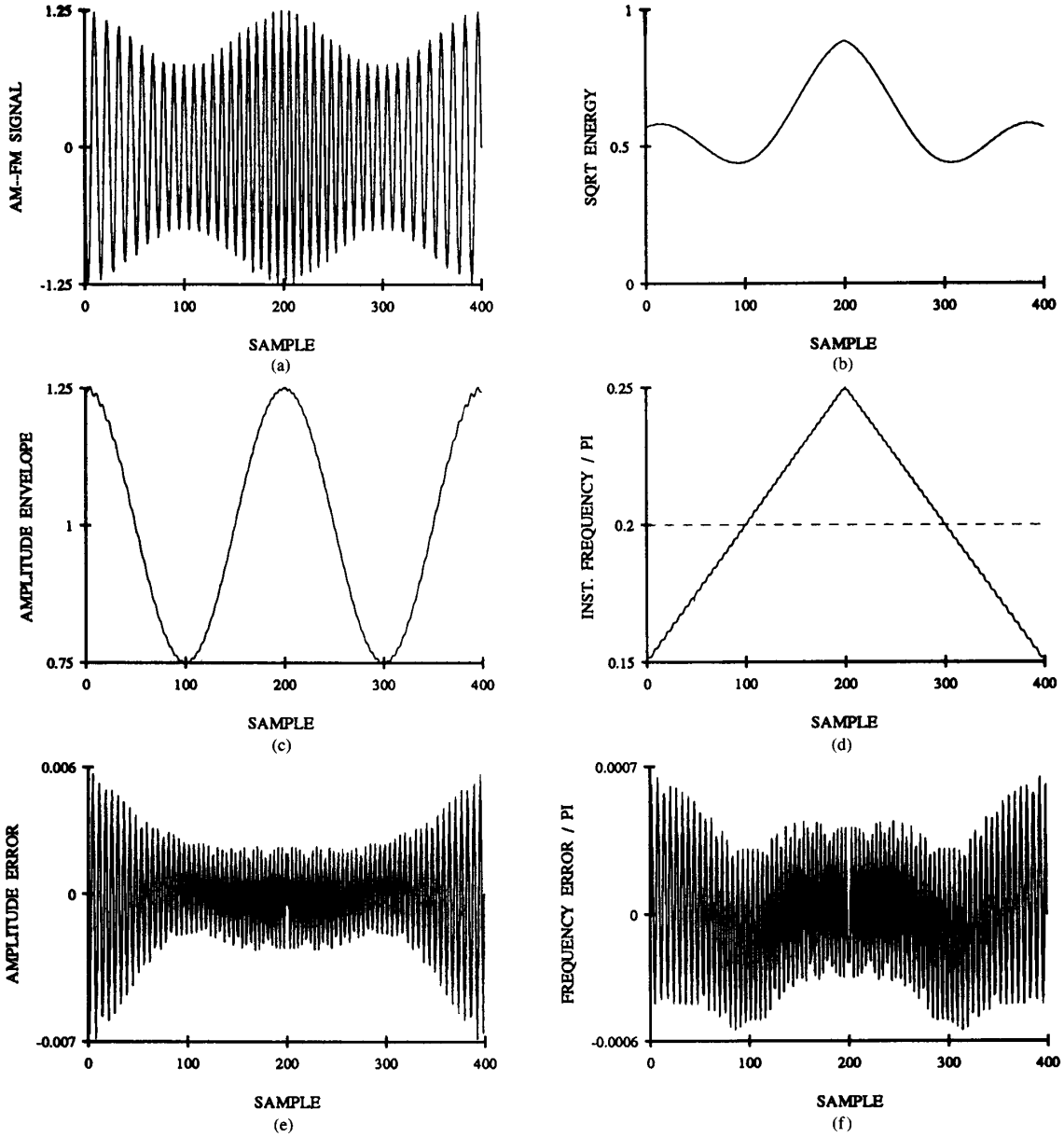


Fig. 2. (a) AM-FM/Linear signal $x(n) = a(n) \cos [0.2\pi n + \pi(n - 100)^2/4000)]$ for $n = 0, \cdots, 200$ and $x(n) = a(n) \cos [0.25\pi n - \pi(n - 200)^2/4000 + \pi/2]$ for $n = 201, \cdots, 400$, where $a(n) = [1 + 0.25 \cos (\pi n/100)]$. (b) $\sqrt{\Psi[x(n)]}$. (c) Estimated amplitude envelope using DESA-1. (d) Estimated instantaneous frequency (shown as fraction of π); the dotted line shows the average value of the true $\Omega_i(n)/\pi$. (e) Error in amplitude estimation. (f) Error in frequency estimation.

$$= \tfrac{1}{2} \arccos \left[1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]} \right] \quad (130)$$

$$|A| = \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}}. \quad (131)$$

Note that if $x(n)$ has been obtained from sampling a continuous-time signal, then the actual frequency f_c (in Hz) can be found from Ω_c and the sampling frequency f_s as $f_c = \Omega_c f_s / 2\pi$.

B. AM-FM Signals

Consider first a discrete-time AM-FM signal $x(n) = a(n) \cos [\phi(n)]$ whose instantaneous frequency signal $\Omega_i(n)$ is a finite sum of cosines as in (78), bandlimited with bandwidth $\Omega_f << 1$ and with maximum deviation $\Omega_m << \Omega_c$, and whose general amplitude signal $a(n)$ is band-

limited with bandwidth Ω_a satisfying (98). Its symmetric difference is

$$s(n) = [x(n+1) - x(n-1)]/2 = D(n) + E(n) \quad (132)$$

where

$$D(n) = a(n) \overbrace{[\cos \phi(n+1) - \cos \phi(n-1)]/2}^{c(n)} \quad (133)$$

$$E(n) = [a(n+1) - a(n)] \cos [\phi(n+1)]/2 + [a(n) - a(n-1)] \cos [\phi(n-1)]/2. \quad (134)$$

Since $\Omega_f << 1$, if follows from (84), (85) that

$$c(n) = -\sin\left[\frac{\phi(n+1) - \phi(n-1)}{2}\right] \cdot \sin\left[\frac{\phi(n+1) + \phi(n-1)}{2}\right] \quad (135)$$

$$\approx -\sin [\Omega_i(n)] \sin [\phi(n)]. \quad (136)$$

Now $D_{\max} \approx a_{\max} \sin (\Omega_i)_{\max}$ and $E_{\max} \approx 2a_{\max} \sin (\Omega_a/2)$. Hence, the order of magnitude of D is much larger than that of E. Thus, ignoring E,

$$s(n) \approx -a(n) \sin [\Omega_i(n)] \sin [\phi(n)]. \quad (137)$$

Since $\Omega_m << \Omega_c$, it follows from the approximation (97) that the amplitude $a(n) \sin (\Omega_i(n))$ of $s(n)$ has an effective bandwidth of $\Omega_a + \Omega_f$. Hence, by (72) and (98),

$$\Psi[s(n)] \approx a^2(n) \sin^4 [\Omega_i(n)]. \quad (138)$$

The above analysis yields the following formulas for estimating the time-varying frequency and amplitude envelope of the AM–FM signal:

$$\frac{1}{2} \arccos\left[1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]}\right] \approx \Omega_i(n) \quad (139)$$

$$\frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}} \approx |a(n)|. \quad (140)$$

We call this the DESA–2 algorithm, where "2" implies the approximation of first-order derivatives by differences between samples whose time indices differ by 2. This DESA uses symmetric differences and thus avoids having to involve values of Ω_i at noninteger time indices. The frequency estimation part assumes that $0 < \Omega_i(n) \leq \pi/2$. Thus, the DESA–2 can be used to estimate instantaneous frequencies $\leq 1/4$ the sampling frequency. This does not present a problem because by doubling the sampling frequency it can be used to estimate frequencies up to $1/2$ the original sampling frequency. Note also that the formula with arccos( ) in (139) can be replaced by an arcsin( ) expression

$$\arcsin\left(\sqrt{\frac{\Psi[x(n+1) - x(n-1)]}{4\Psi[x(n)]}}\right) \approx \Omega_i(n) \quad (141)$$

but this comes at the expense of an additional square-root operation per sample.

We have applied DESA–2 to many cases of AM–FM signals and found that it performs very similarly to the DESA–1. Actually, in most cases it is not possible to see the difference between the DESA–1 and DESA–2 by visual inspection of the resulting signals. Therefore, we resort to numerical comparisons. For each one of the three DESAs, Table I shows the mean absolute and rms values for the amplitude and frequency estimation errors normalized by dividing them with the mean absolute and rms values of the corresponding signals. These results were obtained for the following class of AM–FM/Cosine signals

$$\left[1 + \kappa \cos\left(\frac{\pi}{100} n\right)\right] \cos\left[\frac{\pi}{5} n + 20\lambda \sin\left(\frac{\pi}{100} n\right)\right];$$

$$n = 0, 1, \cdots, 400$$

with instantaneous frequency $\Omega_i(n) = 0.2\pi[1 + \lambda \cos (\pi n/100)]$. The percent errors in Table I are *average* values obtained by computing the corresponding errors for 100 combinations of AM and FM amounts ranging between 5% and 50% at steps of 5%, i.e., for all values $(\kappa, \lambda) \in \{(0.05i, 0.05j)\colon i, j = 1, \cdots, 10\}$. The results in Table I present strong empirical evidence that, for a ratio of information bandwidth versus carrier in the order of $\Omega_a/\Omega_c = \Omega_f/\Omega_c = 1/20$, on the average all three DESA's perform very well with errors smaller than 1%, measured using both the mean absolute and the rms norm. Both DESA–1 and DESA–2 clearly outperform DESA–1a. Also, DESA–1 performs slightly better than DESA–2. The difference in their performance was only in the order of 0.01%–0.1% and may be attributed to the facts that DESA–1 uses a smoothing average of the energy signals from the forward and backward difference and that DESA-2 uses a 3-sample symmetric difference which is a coarser approximation to the first time derivative than the 2-sample differences used by DESA–1. However, as discussed later in Section VI-A, DESA–2 is the fastest of all three DESA's; in addition, its mathematical analysis is the simplest.

Another issue regarding the DESA's is how much their performance deteriorates in the presence of noise. A rigorous treatment of this issue goes beyond the scope of this paper, whose purpose is the development of the basic theory of ESA's for AM–FM signals with application to rel-

TABLE I
PERCENT AMPLITUDE AND FREQUENCY ESTIMATION ERRORS USING DESA'S ON AM-FM/COSINE SIGNALS

Algorithm	Amplitude Estimation mean absolute error %	rms error %	Frequency Estimation mean absolute error %	rms error %
DESA-1a	0.75	0.97	0.82	0.97
DESA-1	0.47	0.57	0.33	0.39
DESA-2	0.53	0.64	0.40	0.47

TABLE II
PERCENT ESTIMATION ERRORS USING DESA-1 ON AM-FM/COSINE SIGNALS WITH NOISE

SNR (dB)	Median Filter	Amplitude Estimation mean absolute error %	rms error %	Frequency Estimation mean absolute error %	rms error %
30	no	6.29	10.56	5.55	9.74
30	yes	2.77	4.26	2.68	4.40
20	no	18.19	29.68	17.98	32.11
20	yes	8.93	12.90	10.31	16.04

atively noise-free speech signals. However, for empirical comparisons we provide in Table II some numerical results that show the performance of DESA-1 applied to the same AM-FM/Cosine signals used for Table I in the presence of added white Gaussian noise at two signal-to-noise ratio (SNR) levels. We have generally observed that the estimation errors of the DESA's usually appear as isolated spikes of large amplitude. Hence, an effective way to improve the estimated amplitude and frequency signals is to post smooth them with a median filter. As Table II shows, the use of a 5-point median post smoothing significantly improves the performance of DESA-1 in white noise. Specifically, the DESA-1 followed by a 5-point median yields amplitude and frequency estimation errors less than 5% when the SNR is 30 dB, whereas when the SNR deteriorates to 20 dB the same system yields errors in the order of 10%.

Now let us apply the DESA-2 to the AM-FM/Linear signal $x(n) = a(n) \cos [\phi(n)]$ with the quadratic phase of (111). First, consider the symmetric difference

$$s(n) = [x(n+1) - x(n-1)]/2$$
$$= a(n) \underbrace{[\cos \phi(n+1) - \cos \phi(n-1)]/2}_{c(n)} + E(n) \tag{142}$$

where $E(n)$ is given by (134). From (112) and (113) it follows that

$$c(n) = -\sin [\Omega_i(n)] \sin \left[\phi(n) + \frac{\Omega_m}{N} \right]. \tag{143}$$

Ignoring the $E(n)$ term, whose order of magnitude is much smaller than that of $a(n)c(n)$, yields

$$s(n) \approx -a(n) \sin [\Omega_i(n)] \sin \left[\phi(n) + \frac{\Omega_m}{N} \right]. \tag{144}$$

Assuming $\Omega_m << \Omega_c$ and using (97) implies that the approximate amplitude $a(n) \sin (\Omega_i(n))$ of $s(n)$ has effective bandwidth equal to $\Omega_{\hat{a}}$. Therefore, assuming (89) and (115),

$$\Psi[s(n)] \approx a^2(n) \sin^4 [\Omega_i(n)]. \tag{145}$$

Hence, we can apply the DESA-2 algorithm, exactly as for the general AM-FM case, to approximately recover the instantaneous frequency and amplitude of the AM-FM/Linear signal.

We have performed extensive numerical comparisons among the three DESA's applied to AM-FM signals with sinusoidal amplitude and linearly-varying frequency in the absence or presence of noise, and the results were very similar to the ones reported in Tables I and II. Namely, in the absence of noise, all DESAs yielded amplitude and frequency estimation errors in the order of 1% or less for AM amounts of 5%-50% and FM amounts of 2.5%-25%. The DESA-1 had the best performance. In the presence of noise with SNR = 30 dB, the DESA-1 followed by post smoothing via a 5-point median filter yielded errors less than 10%.

V. APPLICATION TO SPEECH ANALYSIS

The analysis in the previous sections has established that the ESA's can track well the time-varying amplitude and frequency variations in AM-FM signals. Here we apply them to detect modulations in speech resonances. For all the experiments in this section we use the DESA-1.

A. Evidences for Speech Modulations

By "speech resonances," or "cavity resonators," we loosely refer to the oscillator systems formed by local vocal tract cavities emphasizing certain frequencies and deemphasizing others. There are several experimental and

theoretical evidences for the existence of modulations in speech signals. Most of them are centered around recent ideas of analyzing the dynamics of speech production using concepts from fluid dynamics to study the properties of the speech airflow. Our initial inspiration to consider such issues about speech production has come from Teager's work [30]–[34]. In addition, the current availability of fast computers has made it possible (and an interesting research area in its own right) to investigate fluid dynamics phenomena during speech production through numerical simulations [9], [10], [35] of the nonlinear partial differential equations governing the physics of the speech airflow in the vocal tract. Next we list and briefly discuss some evidences for speech modulations.

1) Separated and Unstable Airflow: Teager concluded from his extensive air velocity measurements in the vocal tract that the airflow is highly separated. Related discussion can be found in [11], [31], [33], [34]. Quoting from Teager [33], "· · · During phonation, airflow in the mouth, and most probably in the rest of the vocal tract, is separated, not isotropic. That is to say, the flow, instead of being stable and uniform across any cross section during a single pitch period, is time varying, concentrated near surfaces, and can switch many times between those surfaces · · · ." One can theoretically predict the separation of speech airflow at cavity inlets by using standard arguments from fluid dynamics [36] about the pressure and velocity fields. For our work on speech modulations, the important implication from these considerations is that the air jet flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses, which affects the frequency of a cavity resonator.[2]

2) Vortices: During speech production vortices can easily build up that can encircle the air jet passing through. These vortices can act as modulators of the energy of the jet. Vortices in the speech airflow have been experimentally found in Teager's work and in numerical simulations of the vocal tract [9], [35]. They have also been theoretically predicted in [22], [32] using simple geometries.

3) Oscillators with Time-Varying Elements: Even in simple second-order oscillators it is known that slow time variations of the oscillator elements can result in amplitude or frequency modulation of the simple oscillator's cosine response. For example, consider an undriven undamped oscillator consisting of a mass m and a spring with stiffness coefficient k. The motion equation is

$$\ddot{x} + \omega_i^2 x = 0, \quad \omega_i^2 = \frac{k}{m} \tag{146}$$

where $x(t)$ is the displacement. If m or k are time varying, then the frequency ω_i is also time-varying. For example, assume it can be modeled as

$$\omega_i^2(t) = \omega_c^2 \left[1 + 2 \frac{\omega_m}{\omega_c} \cos(\omega_f t) \right]. \tag{147}$$

If $\omega_m << \omega$ and $\omega_f << \omega_c$, it has been shown in [3], [37] that the *approximate* solution of (146) is

$$x(t) = A \cos \left[\omega_c t + \frac{\omega_m}{\omega_f} \sin(\omega_f t) \right] \tag{148}$$

which is an FM signal. Similarly second-order oscillators with time-varying damping generate responses that contain amplitude modulation [38]. Thus, during speech production, the time-varying air masses and effective cross-sectional areas of vocal tract cavities that rapidly vary following the separated airflow can cause modulations of the pressure and velocity fields.

4) Energy Pulses: Teager [34] found experimental evidence that speech resonances exhibit modulation structure that cannot originate from a second-order linear resonator model. To see this let us assume that we have a discrete-time second-order linear resonator system with impulse response

$$x(n) = Ar^n \cos(\Omega_c n + \theta). \tag{149}$$

Then applying Ψ to $x(n)$ yields

$$\Psi[Ar^n \cos(\Omega_c n + \theta)] = A^2 r^{2n} \sin^2(\Omega_c). \tag{150}$$

This is illustrated in Figs. 3(a) and (b). In the same figure the DESA is used to estimate the exponentially-decaying amplitude envelope $|Ar^n|$ and the constant frequency Ω_c. Thus, if a signal representing a speech resonance were produced by a second-order linear resonator, then this signal would cause an exponentially-decaying output from the energy operator. In contrast, Teager found that bandpass filtering speech vowel signals around their formants and then applying the energy operator often yielded several pulses, which he called "*energy pulses*," per pitch period. These energy pulses indicate some kind of modulation in each formant. In [30], [32] it appears that Teager also did some work, which he called "microdissection," on trying to isolate the instantaneous amplitude and frequency of such modulations, although the mathematical details and algorithms used are not described therein. In our speech experiments [17], [18] we also found these energy pulses and attempted to model them using AM–FM signals, as explained next.

B. Speech Resonances

All the previous discussion paved the way and motivates now the modeling of a single speech resonance (in discrete time) by an exponentially-damped AM–FM

[2]The time varyingness of speech formant frequencies caused by the modulation phenomena discussed in this paper occur at time scales much smaller than the pitch period; thus they are a microtime scale phenomenon. Therefore, they should be distinguished from formant variations caused by vocal tract movements, which usually occur at larger time scales of about 10 msec [29] and have been well studied in acoustic phonetics with many speech processing applications as for example recently in [4], [14].

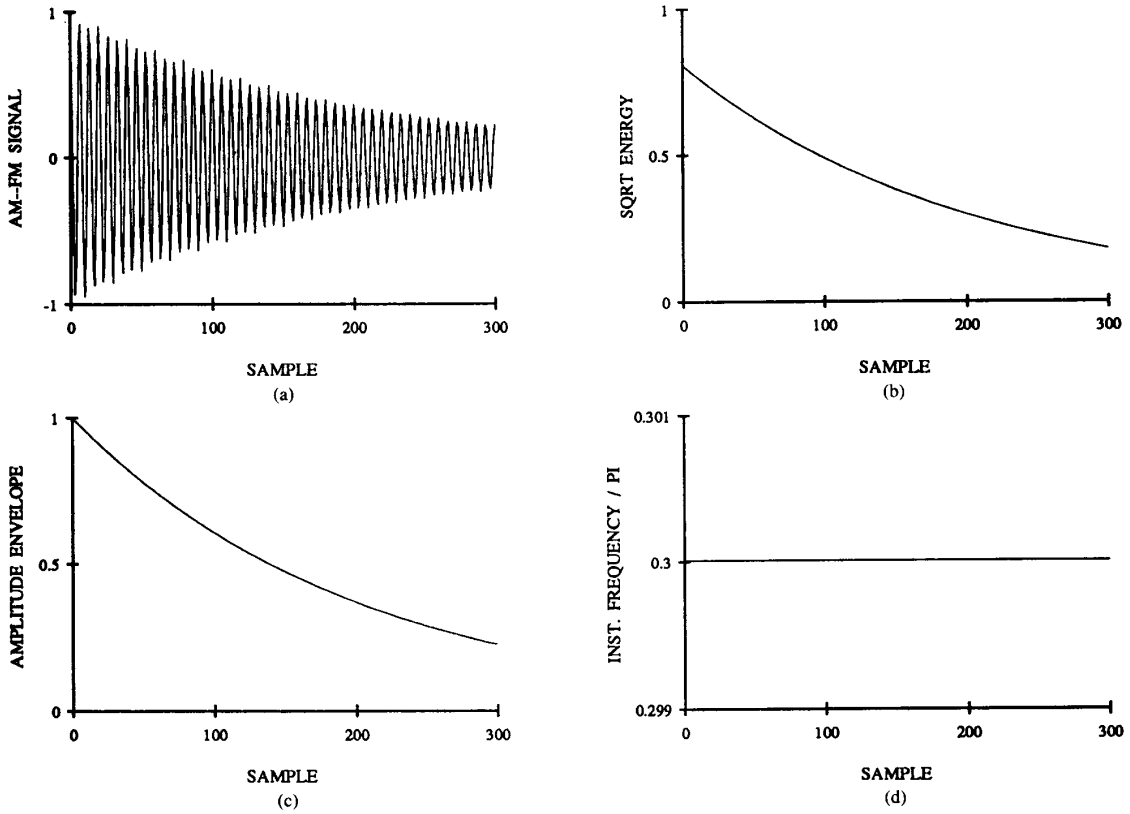


Fig. 3. (a) Exponentially-damped cosine signal $x(n) = (0.995)^n \cos(0.3\pi n)$. (b) $\sqrt{\Psi[x(n)]}$. (c) Estimated amplitude envelope using DESA-1. (d) Estimated instantaneous frequency.

model[3]

$$R(n) = a(n)\cos[\phi(n)] \tag{151}$$

$$= r^n A(n)\cos\left[\Omega_c n + \Omega_m \int_0^n q(k)\,dk + \phi(0)\right] \tag{152}$$

where Ω_c is the center frequency value of the formant, the instantaneous frequency $\Omega_i(n) = \Omega_c + \Omega_m q(n)$ models the time-varying formant whose deviation from Ω_c follows some modulating signal $q(n)$, $|q(n)| \leq 1$, $A(n)$ is some time-varying amplitude, and $r \in (0, 1)$ is related to the rate of energy dissipation. By assuming that the total amplitude $a(n)$ and instantaneous frequency $\Omega_i(n)$ do not vary too fast in time or too greatly compared with the carrier Ω_c, it follows from the discussion in Sections III and IV that

$$\sqrt{\Psi[R(n)]} \approx r^n |A(n)\sin(\Omega_c + \Omega_m q(n))|. \tag{153}$$

Thus, $\sqrt{\Psi[R(n)]}$ is an exponentially-damped product of the envelope and the sine of the instantaneous frequency of the resonance. For notational simplicity we will henceforth hide the decay factor r^n into the total amplitude signal $a(n) = r^n A(n)$; this causes no loss of generality because $\Psi[r^n x(n)] = r^{2n}\Psi[x(n)]$ for any signal x.

An example of an exponentially-damped AM-FM input signal is shown in Fig. 4(a). Fig. 4(b) shows the $\sqrt{\Psi}$'s output, which is approximately equal to the product of the amplitude envelope (AM component) and the sine of the instantaneous frequency (FM component). However, the AM component dominates and visually hides the FM component, because the amount of AM is much larger than the amount of FM. The relative error magnitude in

[3]There has been some previous work in modeling nonstationarity in speech signals. In [21] the speech signal was modeled as a sum of harmonic sine-wave components (not resonances) with time-varying amplitudes and frequencies. However, during the estimation part in [21] constant amplitudes and frequencies were assumed over an analysis frame; during the synthesis part time-varying quadratic frequencies and linear amplitudes were allowed. In another work [20] nonstationarity in speech signals was modeled by using continuous-time sinusoids with time-varying frequency and amplitude. However, the amplitude was constrained to be only a Gaussian, exponential, or constant function, whereas the instantaneous frequency was constrained to only vary linearly. In contrast, our approach allows for arbitrary amplitude and frequency signals, that are either bandlimited or slow varying. Also, the above works model speech harmonics whereas our work deals with resonances. Finally, in [20] the parameters of the sinusoids were found by minimizing a weighted average squared spectral error criterion based on short-time Fourier transforms. In contrast, our approach using energy operators and ESA's is more efficient due to the simplicity of these nonlinear operators and can adapt much more rapidly to speech nonstationarities due to the almost instantaneous nature of the energy operators.

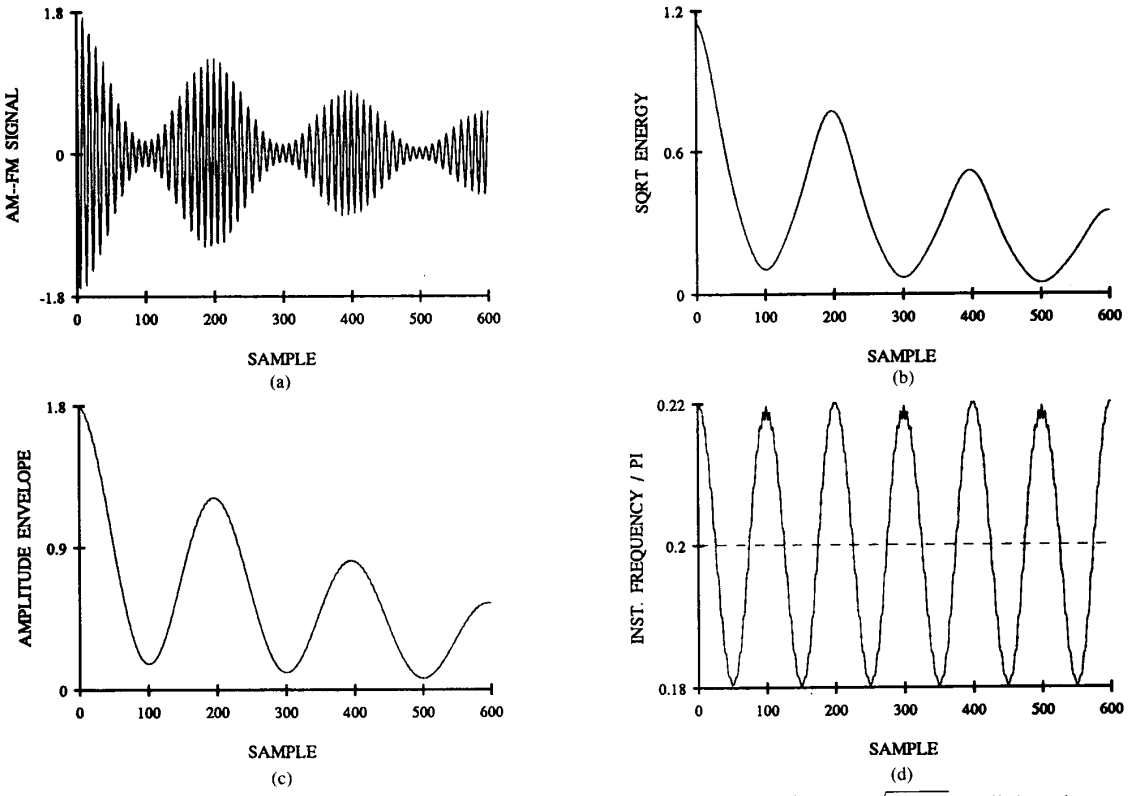


Fig. 4. (a) AM–FM signal $x(n) = (0.998)^n[1 + 0.8 \cos(\pi n/100)] \cos[\pi n/5 + \sin(\pi n/50)]$. (b) $\sqrt{\Psi[x(n)]}$. (c) Estimated amplitude envelope using DESA-1. (d) Estimated instantaneous frequency.

$\sqrt{\Psi}$ estimating the product of the AM and FM components was 0.2%. Despite the mixing of the two components in $\sqrt{\Psi}$'s output, the DESA was able to successfully separate and estimate the amplitude envelope and instantaneous frequency as shown in Figs. 4(c) and (d).

We believe that the class of exponentially-damped AM–FM signals (152) where $a(n)$ and $\Omega_i(n)$ are narrowband signals that do not vary too fast or too much in time compared to the carrier (e.g., modelable as sums of a few slow-varying sinusoids within a pitch period) may serve as a good model for speech resonances for at least three reasons: i) The shape of the energy pulses in $\sqrt{\Psi}$'s output when the input is a synthetic exponentially-damped AM–FM signal with sinusoidal modulating signals matches well the shape of the energy pulses observed on actual bandpass filtered speech waveforms. This can be seen by comparing the energy pulses in the synthetic example of Fig. 4 and the energy pulses on real speech resonances shown later. ii) As we have already discussed, there exist strong experimental and theoretical evidences for speech resonances to have time-varying formants and amplitudes. iii) The modulating signals in AM–FM models can be efficiently estimated by using the DESA's.

Experiments: Here we describe some of our experiments on detecting modulations in speech resonances. We extract a single resonance by bandpass filtering the speech with a Gabor filter [7], whose impulse and frequency response are

$$h(t) = \exp(-\alpha^2 t^2) \cos(\omega_c t) \tag{154}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2\alpha}\left(\exp\left[-\frac{(\omega - \omega_c)^2}{4\alpha^2}\right] + \exp\left[-\frac{(\omega + \omega_c)^2}{4\alpha^2}\right]\right). \tag{155}$$

The reasons for selecting the above bandpass filter are twofold: i) It is optimally compact in the time and frequency domains, because its rms time and frequency width product assumes the minimum value in the uncertainty principle inequality; ii) The Gaussian shape of $H(\omega)$ avoids producing sidelobes (or big sidelobes after truncation of h) that could produce false pulses in the Ψ's output.

Our design of the discrete bandpass Gabor filter proceeds as follows: A center formant frequency f_c is selected from the short-time speech spectrum.[4] A value of α is

[4]In this paper we simply position the bandpass filters centered around manually-found formant spectral peaks. However, *automated* adjustment of the filter center frequencies to approximately coincide with the formant center frequencies is possible as follows: It has been found in [8] that iteratively applying the Gabor bandpass filter (initially centered anywhere in the vicinity of a formant spectral peak) and the DESA by updating the center filter frequency as the average of the estimated instantaneous frequency converges to the true center formant/filter frequency after a few iterations.

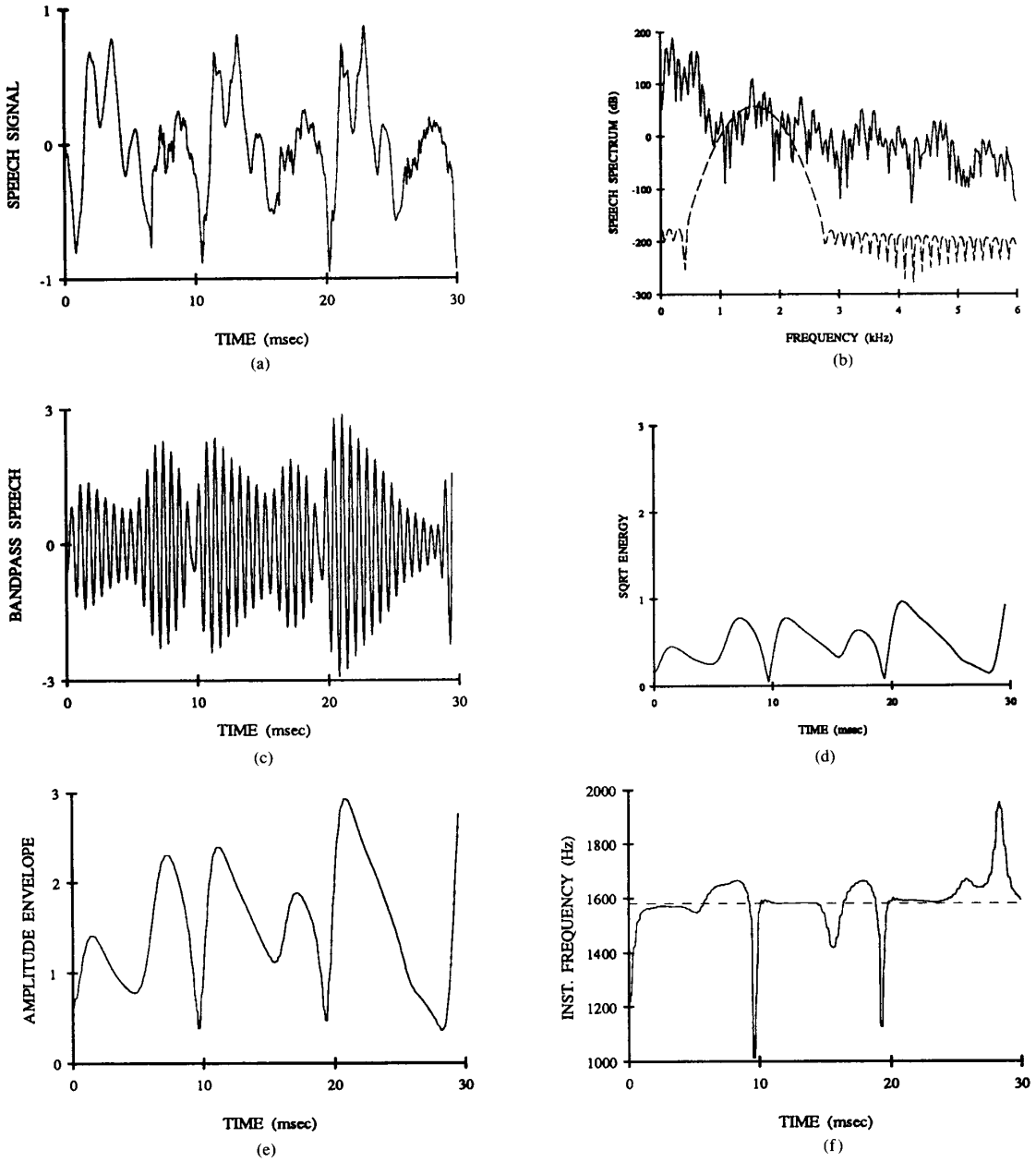


Fig. 5. (a) Signal $s(n)$ from speech vowel /E/ sampled at 30 kHz. (b) Spectral magnitude (magnified by 5) of $s(n)$. Dotted line shows the magnitude response of the Gabor bandpass filter around f_c = 1580 Hz (α = 1000, N = 102). (c) Signal $x(n)$ from Gabor bandpass filtering of $s(n)$. (d) $\sqrt{\Psi[x(n)]}$. (e) Estimated amplitude envelope using DESA-1. (f) Estimated instantaneous frequency, smoothed by an 11-point median filter. (The dotted line shows the center formant value.).

selected to control the bandwidth using the rule [7] that the rms bandwidth of the Gabor filter is equal to $\alpha/\sqrt{2\pi}$. $h(t)$ is discretized by replacing t with nT, where T is the sampling period, and truncating $h(n)$ to a symmetric FIR filter $h(n) = \exp(-b^2n^2)\cos(\Omega_c n)$, with $-N \leq n \leq N$, $b = \alpha T$, and $\Omega_c = 2\pi f_c T$. Then the Gabor bandpass filtering is performed by convolving the truncated $h(n)$ with the speech signal. The integer N is chosen to truncate the Gaussian envelope of $h(n)$ essentially to zero; e.g., $\exp(-b^2N^2) = 10^{-5}$.

Fig. 5 shows (a) a 30 msec segment of a speech vowel /E/ from the word "f/ea/ther" sampled at f_s = 30 kHz; (b) its spectral magnitude superimposed with the magnitude response of a bandpass Gabor filter centered at a formant f_c = 1580 Hz; (c) the speech signal part $x(n)$ corresponding to this frequency band; (d) the energy operator

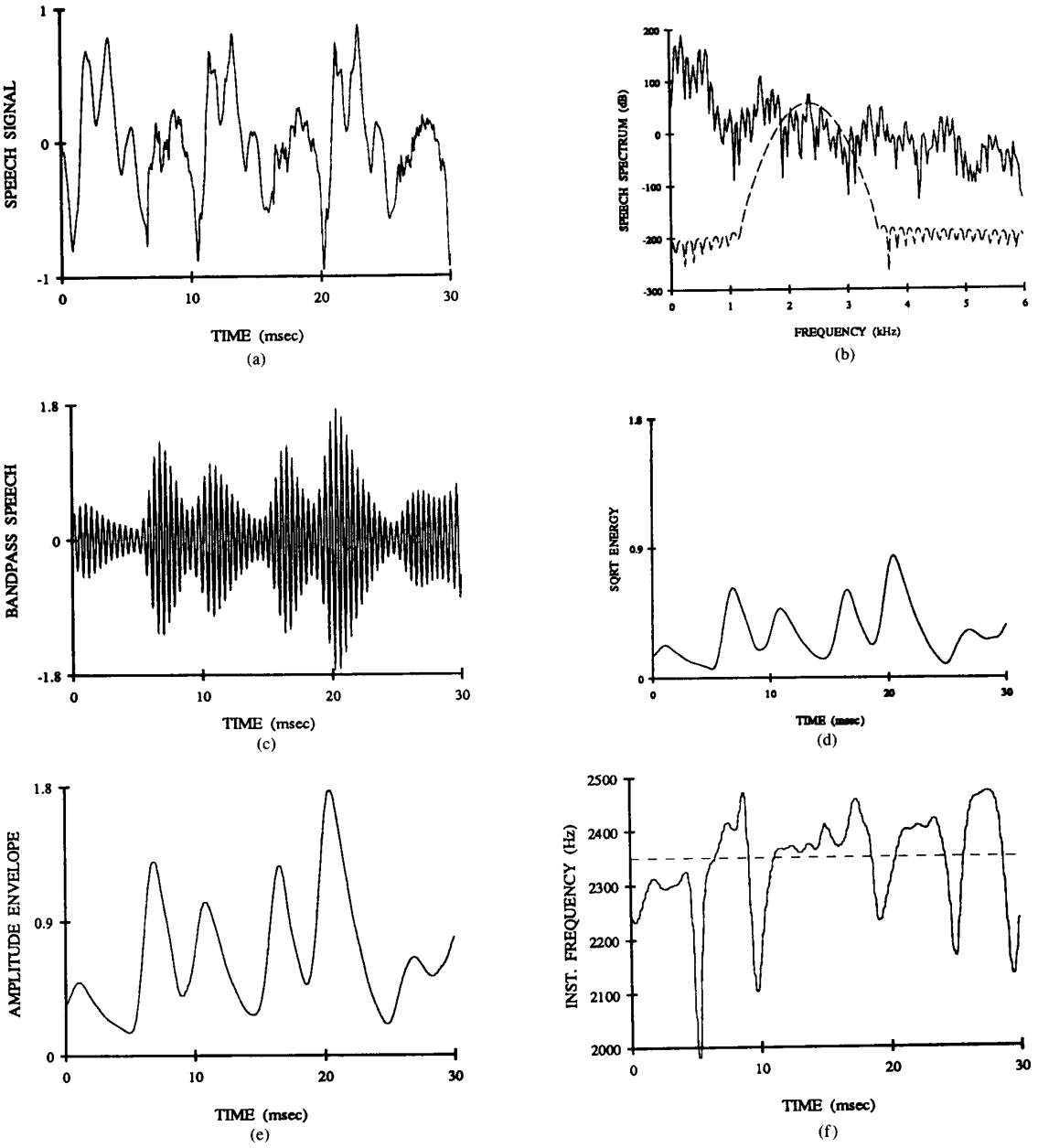


Fig. 6. (a) Signal $s(n)$ from speech vowel /E/ sampled at 30 kHz. (b) Spectral magnitude (magnified by 5) of $s(n)$. Dotted line shows the magnitude response of the Gabor bandpass filter around $f_c = 2350$ Hz ($\alpha = 1000$, $N = 102$). (c) Signal $x(n)$ from Gabor bandpass filtering of $s(n)$. (d) $\sqrt{\Psi[x(n)]}$. (e) Estimated amplitude envelope using DESA-1. (f) Estimated instantaneous frequency, smoothed by an 11-point median filter. (The dotted line shows the center formant value.)

applied on $x(n)$; (e) the amplitude envelope and (f) the instantaneous frequency extracted using the DESA. The frequency signal was post-smoothed by a median filter to suppress a few isolated impulse-like spikes occurring at amplitude valleys. The experiment reported in Fig. 5 is repeated for the same speech vowel segment but for its next two higher formants, i.e., 2350 Hz in Fig. 6 and 3400 Hz in Fig. 7. As Figs. 5(d), 6(d), and 7(d) show, there are present two (2–3) *energy pulses* per pitch period. We have seen similar numbers of energy pulses in many other of our experiments with signals from speech vowels. As explained before, these multiple energy pulses per pitch period in the output of Ψ applied on a speech resonance signal indicate the existence of modulations. Namely, if the resonance were linear with constant amplitude and frequency, then the energy operator's output

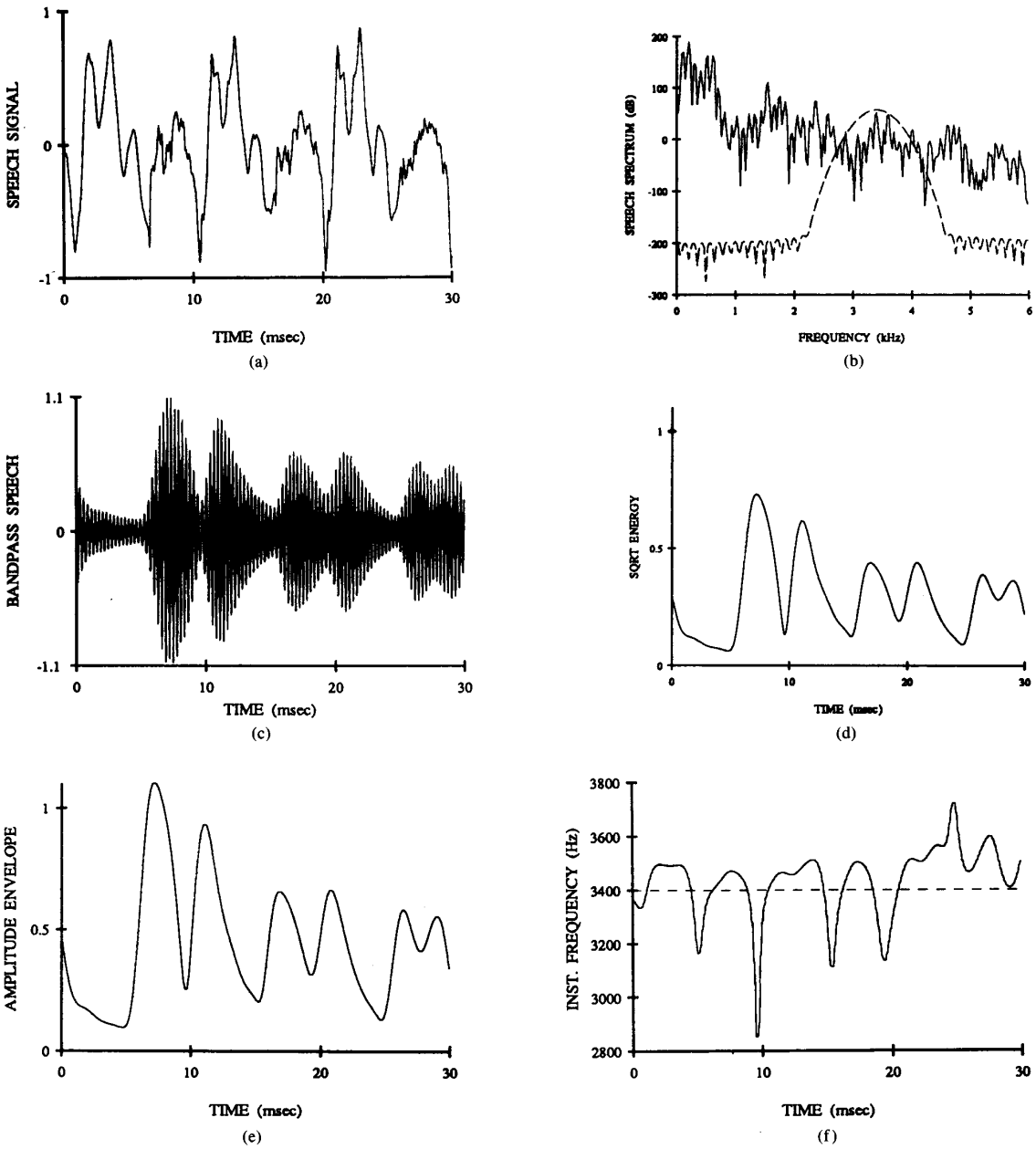


Fig. 7. (a) Signal $s(n)$ from speech vowel /E/ sampled at 30 kHz. (b) Spectral magnitude (magnified by 5) of $s(n)$. Dotted line shows the magnitude response of the Gabor bandpass filter around $f_c = 3400$ Hz ($\alpha = 1000$, $N = 102$). (c) Signal $x(n)$ from Gabor bandpass filtering of $s(n)$. (d) $\sqrt{\Psi[x(n)]}$. (e) Estimated amplitude envelope using DESA-1. (f) Estimated instantaneous frequency, smoothed by an 11-point median filter. (The dotted line shows the center formant value.)

would be a damped exponential with no pulses as in (150). The structure of the damped AM–FM signal (152) with sinusoidal AM and FM modulating signals (or superpositions of a few sinusoids) and its output (153) from Ψ may approximately explain the shape of these measured energy pulses. This is further supported by the shape of the actual amplitude envelope and instantaneous frequency that the separation algorithm has extracted from the speech resonances.

As Figs. 5(e), (f), 6(e), (f), and 7(e), (f) show, the energy separation algorithm uncovers some interesting modulation structures in speech resonances. We see there very strong AM modulation; further the amplitude envelope seems to closely follow the energy operator's output. Thus the AM component seems to visually dominate the FM component in the energy operator's output. Nevertheless, the FM is strong especially in the two higher formants. Specifically, the estimated instantaneous fre-

quency signals seem to oscillate around their center formant value with the frequency deviations whose local maxima can reach 200 Hz. However, the instantaneous variation of the lower formant is sometimes small, as shown in the first part of Fig. 5(f) where it increases in the second part of the speech segment. In higher formants we have generally observed stronger modulations, e.g., see Fig. 7(f). We have also found that the instantaneous frequency profiles estimated via the DESA often contain isolated narrow spikes, which are usually caused either by amplitude valleys or by the onset of a new pitch pulse; the latter issue is further explained in a subsequent section. At valleys of the amplitude envelope, the energy signal goes close to zero, thereby creating abrupt spikes in the frequency profile. Excluding these narrow spikes, in vowels the instantaneous frequency and amplitude envelope profiles follow a roughly sinusoidal pattern. Note that the signals used in Figs. 5–7 were sampled at 30 kHz. The reason for using this rather high sampling frequency was to be able to observe the variations of the amplitude and frequency signals at finer time scales. We have observed very similar patterns for a 10 kHz sampling frequency. Finally, there have been a few cases of signals from low formants of vowels where we observed only one major energy pulse per pitch period. This may be partially explained by a large amount of damping, or by a low amount of AM-FM modulation for the specific speaker/sound/formant combination.

Fig. 8 shows a similar speech experiment but for the voiced fricative /Z/ and for a formant around $f_c = 6300$ Hz. We observe there that the instantaneous frequency and amplitude envelope profiles have more complicated shapes than the amplitude/frequency profiles observed in vowels. This is perhaps due to the random bursts of energy that accompany frication. Our understanding of the modulation occurring during frication is still incomplete. As a preliminary model, we speculate that a resonance of a fricative speech signal can be modeled as an AM–FM signal whose modulating signals $a(n)$ and $q(n)$ are mostly random.

Note that our modeling of a *single* resonance in the vocal tract output using an AM–FM signal does not explicitly take into consideration the facts that actual speech vowels are quasi-periodic and usually consist of multiple resonances. Both of these phenomena introduce an additive component to the single resonance which may alter the output of the energy operator and the DESA estimates. Next we briefly analyze these two issues.

C. Effects of Pitch

The jumps in the output of Ψ in the vicinity of the onset of a new pitch period can be explained by observing that the input single-resonance signal in the neighborhood of a pitch impulse can be modeled as the product of an AM–FM signal $x(n) = a(n) \cos[\phi(n)]$ and a unit step signal $u(n)$, where the time origin $n = 0$ has been taken at the location of some pitch "impulse." The output of Ψ to such a signal is

$$\Psi[x(n)u(n)] = \Psi[x(n)]u(n) + \delta(n) \cdot [x^2(0) - \Psi(x(n))_{n=0}] \quad (156)$$

since $\Psi[u(n)] = \delta(n)$ where δ is the discrete-time unit impulse. Hence the output of Ψ contains a right-sided version of $\Psi[x(n)] \approx a^2(n) \sin^2[\Omega_i(n)]$ and an impulse at the location of the pitch impulse. These impulse jumps in the output of Ψ due to the pitch periodicity account for some of the observed jumps in the amplitude envelope and the spikes in the instantaneous frequency estimated from applying the DESA to the speech resonance signal.

To further understand the effects of pitch, we next examine the structure of the amplitude and frequency signals obtained by applying the DESA to resonances from simple vowel-like signals synthesized using time-invariant linear resonators. The motivation is to compare them with our experimental results from real speech resonances. First note from Fig. 3 that a signal produced by a second-order linear resonator with no pitch periodicity yields an exponentially-decaying amplitude envelope and a constant frequency. If we add the pitch periodicity, e.g., by exciting the above linear resonator with a periodic impulse train of period 100 samples, then we obtain the signal in Fig. 9. This signal mimics the structure of a synthetic vowel with a single constant formant at 1500 Hz and a pitch frequency of 100 Hz. (The sampling frequency was set equal to 10 kHz). Then Fig. 9 shows that the $\sqrt{\Psi}$ output and the estimated amplitude envelope via the DESA consist of exponentially-decaying segments interrupted by discontinuity jumps at the time locations of the pitch impulses. The estimated frequency signal is everywhere roughly constant, as it should be, except at the location of pitch impulses where it has large spikes. Finally, Fig. 10 shows a segment from a synthetic vowel generated from the parallel superposition of two single-formant vowel synthesizers of the same type as used in Fig. 9. This synthetic signal has two formants, one at 500 Hz with relative gain 1 and the other at 1500 Hz with relative gain 0.5. The rest of Fig. 10 reports the same experiment as in the real speech experiments of using a Gabor bandpass filter to extract the formant around 1500 Hz and then using the energy operator and the DESA to estimate the instantaneous amplitude and frequency of this resonance. As expected, the amplitude signal consists of exponential-decaying segments with jumps at pitch impulse locations, whereas the frequency signal is everywhere constant except for some narrow doublet-like pulses around the pitch impulse locations. These pitch-induced doublets have a very small height of 20 Hz and are blurred counterparts of the spikes in Fig. 9.

Concluding, the DESA applied to signals corresponding to resonances of vowels synthesized using linear time-invariant resonators yields exponentially-decaying amplitude envelopes and constant formant frequencies except for some narrow pitch-induced spikes. In contrast, the DESA has uncovered energy pulses in the amplitude and

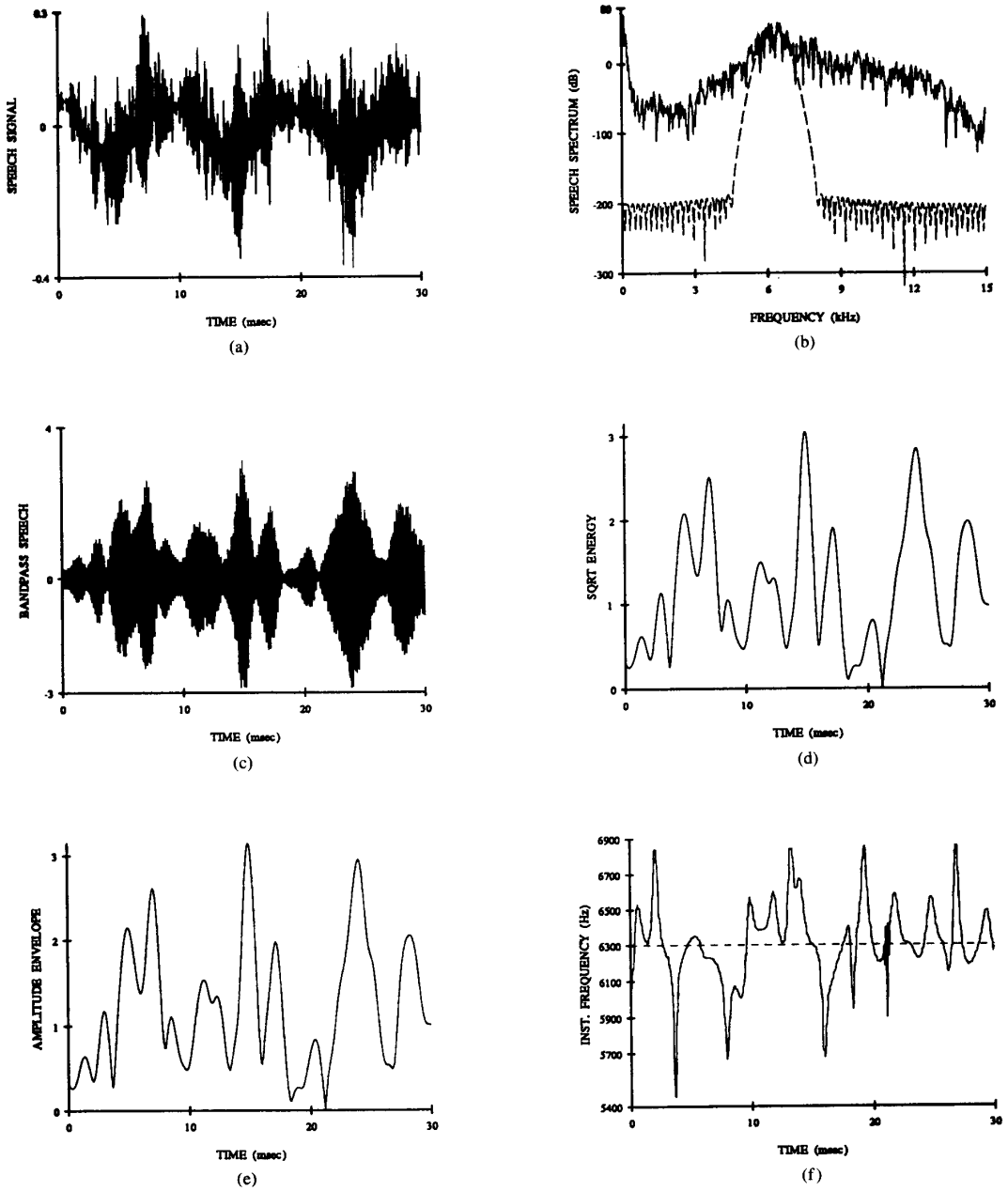


Fig. 8. (a) Speech signal $s(n)$ from fricative /Z/ sampled at 30 kHz. (b) Spectral magnitude (magnified by 3) of $s(n)$. Dotted line shows the Gabor bandpass frequency response around $f_c = 6300$ Hz ($\alpha = 1500$, $N = 68$). (c) Signal $x(n)$ from Gabor bandpass filtering of $s(n)$. (d) $\sqrt{\Psi}[x(n)]$. (e) Estimated amplitude envelope. (f) Estimated instantaneous frequency, smoothed by a 9-point median filter. (The dotted line shows the center formant value.)

frequency signals of real speech resonances, which indicates the existence of modulations in real speech signals.

D. Effects of Neighboring Formants

Consider the case that arises when, inside the passband of the bandpass filter used to extract a speech resonance, there are two formants closely spaced with approximately equal gains. Let us model this situation with the signal

$$x(t) = \sin(\omega_1 t + 2\theta) + \sin(\omega_2 t)$$
$$= 2\cos(\omega_a t + \theta)\sin(\omega_c t + \theta) \quad (157)$$

where ω_1, ω_2 are the formant center frequencies and $\omega_1 \leq \omega_2$. Then $x(t)$ is an AM signal whose carrier and envelope frequencies are $\omega_c = (\omega_1 + \omega_2)/2$ and $\omega_a = (\omega_2 - \omega_1)/2$,

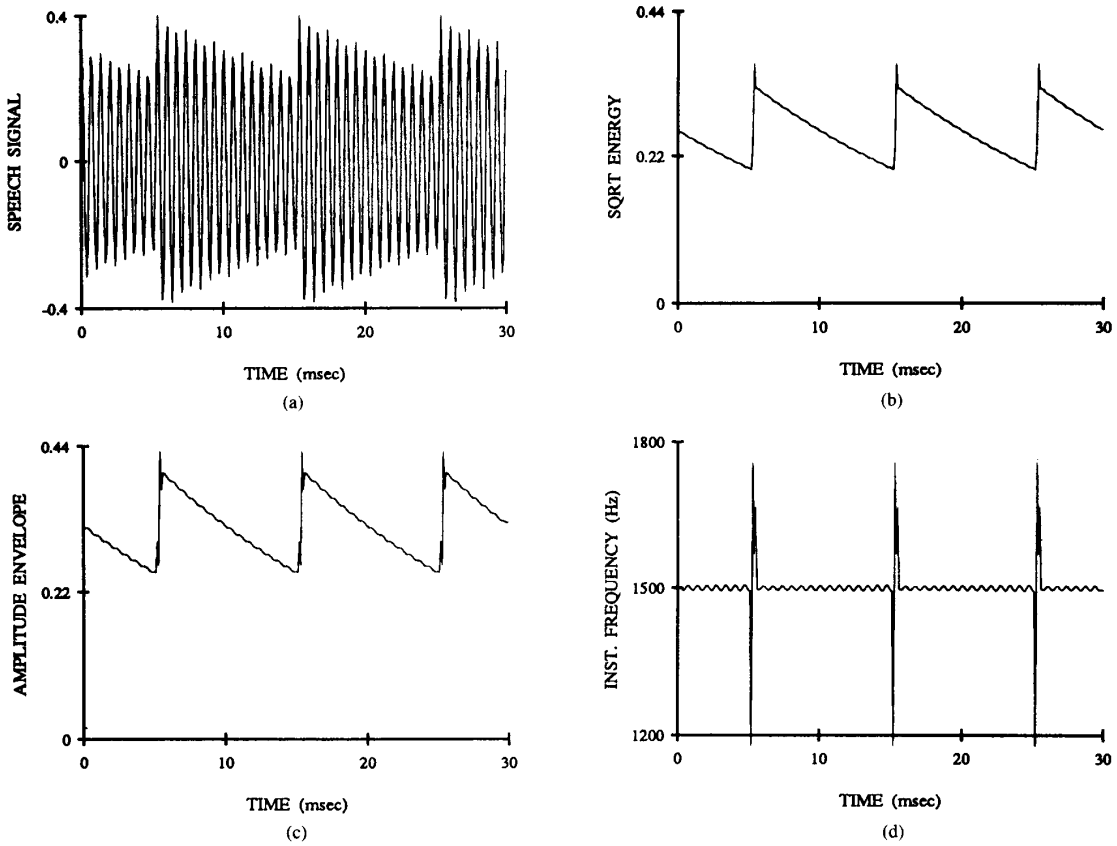


Fig. 9. (a) Signal $x(n)$ from synthetic speech vowel with one formant at 1500 Hz and pitch frequency of 100 Hz. (b) $\sqrt{\Psi[x(n)]}$. (c) Estimated amplitude envelope using DESA-1. (d) Estimated instantaneous frequency.

respectively. Thus, $\sqrt{\Psi[x(t)]} \approx 2\omega_c|\cos(\omega_a t + \theta)|$, and hence $\sqrt{\Psi}$ will track the envelope, if the relative approximation error [19] $(1-d)^2/(1+d)^2$ is $\ll 1$, where $d = \omega_1/\omega_2 \leq 1$. For this error to be $< 10\%$, we must have $d > 0.5$, i.e., the two formants must be less than an octave apart. Then, we observe an AM modulation of one formant by the other. Consider also the case of two consecutive pitch harmonics falling within the resonance bandwidth and passing through the bandpass filter. Then, the above two-sine model holds and may predict a possible tracking of a pitch-related AM envelope. However, this envelope varies with a frequency roughly equal to the pitch frequency and thus the modulation does not introduce additional pulses over a pitch period. In addition, we have experimentally observed in voiced speech resonances that the estimated instantaneous frequency may contain "parasitic" small-magnitude ($\pm 10 - 30$ Hz) ripples due either to the pitch harmonics or to neighbor formants if the bandpass filter has not completely rejected the neighbor formants. A preliminary model we have thought for this case could be a sum of two cosines, one at the center resonance frequency ω_c and another at some neighbor frequency ω_x with much smaller amplitude by a factor $\beta \ll 1$ (to model a residual frequency peak not completely rejected by the filter):

$$\begin{aligned} y(t) &= \cos(\omega_c t) + \beta\cos(\omega_x t + \theta) \\ &\approx \underbrace{\cos[\omega_c t - \beta\sin(\omega_f t - \theta)]}_{\text{FM}} \\ &\quad + \underbrace{\beta\cos(\omega_f t - \theta)\cos(\omega_c t)}_{\text{AM}} \end{aligned} \tag{158}$$

where $\omega_f = \omega_c - \omega_x$. Thus, imperfect bandpass filtering may introduce a parasitic FM component with a very small modulation index $\beta \ll 1$ whose modulating frequency is the difference between the central and the neighbor frequency peak and the carrier is the central frequency. The parasitic AM component has a smaller (by β) order of magnitude than the FM and hence can be ignored. If the neighbor peak is a pitch harmonic, then ω_f will be the pitch frequency and hence this parasitic FM will be very slow varying. The maximum frequency deviation in the estimated instantaneous frequency using an energy separation algorithm will be approximately equal to $\beta\omega_f$; it is this product that determines the peak ripple of the observed parasitic FM component. Preliminary experiments

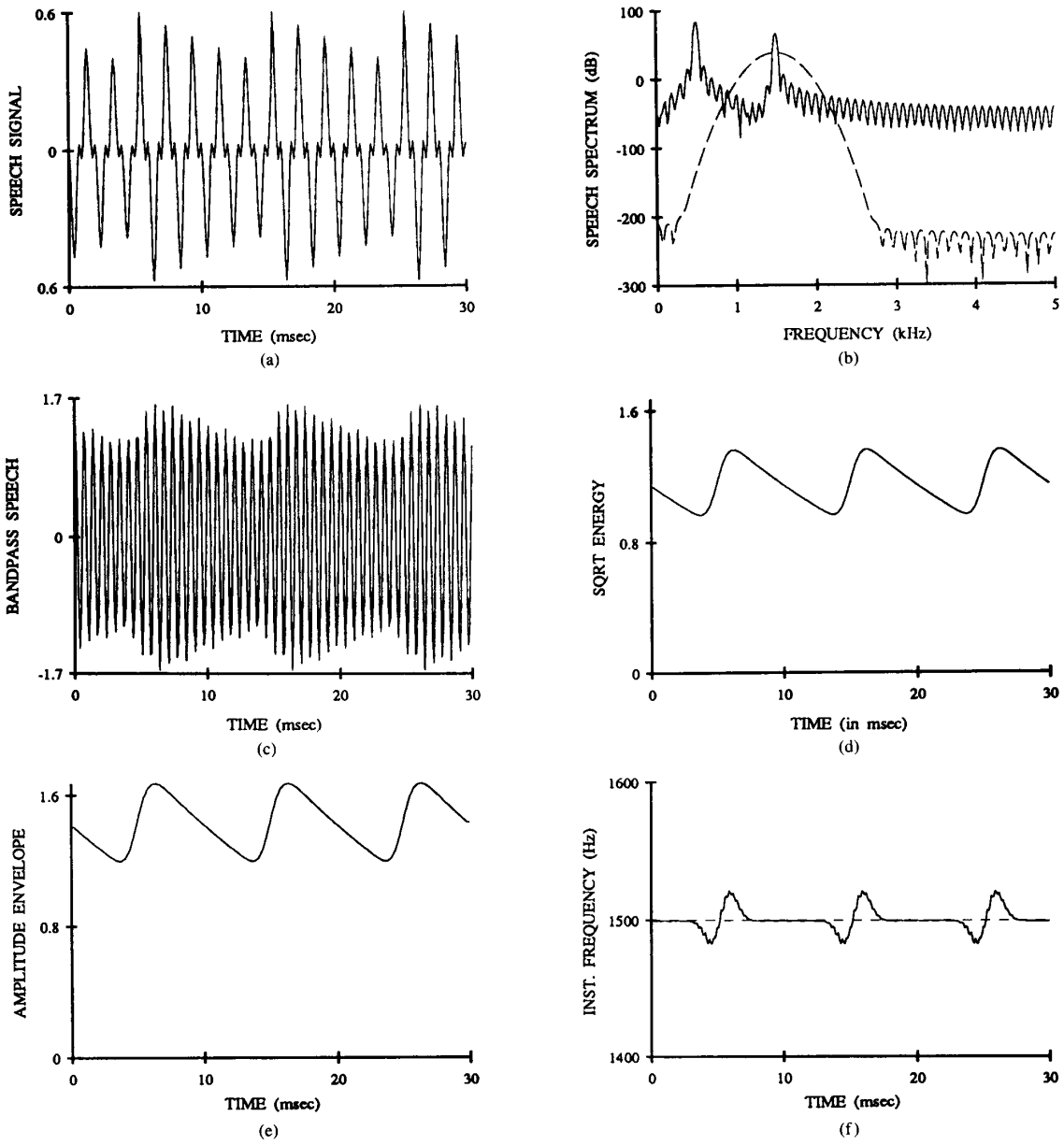


Fig. 10. (a) Signal $x(n)$ from a synthetic speech vowel with two formants at 500 and 1500 Hz and pitch frequency of 100 Hz. (b) Spectral magnitude (magnified by 3) of $s(n)$. Dotted line shows the magnitude response of the Gabor bandpass filter around $f_c = 1500$ Hz ($\alpha = 1000$, $N = 34$). (c) Signal $x(n)$ from Gabor bandpass filtering of $s(n)$. (d) $\sqrt{\Psi[x(n)]}$. (e) Estimated amplitude envelope using DESA-1. (f) Estimated instantaneous frequency.

confirm the above implications of this conjectured model for parasitic FM.

The above discussion implies that the passband of the bandpass filters extracting speech resonances should not be too wide because then they will include significant contributions from neighboring formants which may cause parasitic modulations. On the other hand, the bandpass filters should not have a very narrow passband because this would miss or deemphasize some of the modulations. Optimal decision schemes for choosing the filter bandwidth are presently being investigated.

VI. Discussion

In this section we conclude by briefly discussing a variety of issues related to the ESA's and their application to speech analysis.

TABLE III
COMPUTATIONAL COMPLEXITY OF DESA'S
(NUMBER OF OPERATIONS PER SAMPLE)

Algorithm	Additions	Multiplications	$\arccos(\cdot)$	$\sqrt{(\cdot)}$	W
En. Op. $\Psi(\cdot)$	1	2	0	0	3
DESA-1a	5	8	1	1	4
DESA-1	6	8	1	1	5
DESA-2	4	8	1	1	5

A. Computational Complexity of DESA's

The DESAs are very simple algorithms to implement with almost trivial computational complexity. Table III compares the complexity of the three DESAs, by assuming that additions and subtractions have the same complexity, and so do multiplications and divisions. Thus all DESAs have linear complexity $O(N)$ where N is the length of the analyzed signal segment in samples and the constants in this $O(N)$ are very small. DESA-2 is the fastest and DESA-1 is the slowest, but the differences are very small. In addition, note that all three DESA's behave almost as instantaneous signal operators because they are defined based on an extremely short moving window. Specifically, DESA-1 and DESA-2 involve only a moving window of five samples (at times n, $n \pm 1$, $n \pm 2$), whereas DESA-1a involves a 4-sample moving window (at times n, $n \pm 1$, $n - 2$).

B. Alternative Amplitude/Frequency Separation Algorithms

In addition to the ESA's that we developed in this paper, it is possible to find other approaches to estimate the amplitude and frequency components of an AM-FM signal. Next, we comment on three such approaches.

1) Continuous-Time Energy Separation: For a cosine

$$x(t) = A \cos(\omega_c t + \theta)$$

the following formulas were provided in [39, p. 159] (and in [28] only for the frequency) to exactly compute its constant amplitude A and frequency ω_c:

$$\omega_c^2 = -\frac{\ddot{x}(t)}{x(t)} \tag{159}$$

$$A^2 = x^2(t) - \frac{x(t)\dot{x}^2(t)}{\ddot{x}(t)}. \tag{160}$$

We note that the above algorithm contains implicitly the energy operator only in the amplitude estimator; i.e., (160) can be written as

$$A^2 = x^2 - \frac{x\dot{x}^2}{\ddot{x}} = \frac{\Psi(x)}{(-\ddot{x}/x)} = \frac{\Psi(x)}{\omega_c^2}. \tag{161}$$

Comparing the amplitude and frequency estimation equations of our CESA with the above algorithm, we note a fundamental and important difference: In the CESA of (13) and (14) both estimation equations are the quotients of two functions each of which is approximately a low-bandwidth function, whereas in the above algorithm these functions are the signals and their derivatives themselves and hence are not low-bandwidth but rapidly-changing functions. Thus we would expect the CESA to give more stable and less noisy output estimates than the above algorithm.

2) Discrete-Time Energy Separation: In addition to the three DESA's we discussed previously, it is possible to develop alternative algorithms by combining various shifted versions of the signal and their outputs from Ψ to obtain a set of equations whose solution yields estimates of the amplitude and frequency signals. Next we provide one such algorithm, starting from the case of a cosine

$$x(n) = A \cos(\Omega_c n + \theta)$$

with constant amplitude and frequency. Note that

$$\begin{aligned} x(n+1) + x(n-1) &= 2A \cos(\Omega_c) \cos(\Omega_c n + \theta) \\ &= 2 \cos(\Omega_c) x(n) \end{aligned} \tag{162}$$

Hence, the constants Ω_c and $|A|$ can be found from the formulas

$$\Omega_c = \arccos\left[\frac{x(n+1) + x(n-1)}{2x(n)}\right] \tag{163}$$

$$|A| = \sqrt{\frac{\Psi[x(n)]}{1 - \cos^2(\Omega_c)}} = \sqrt{\frac{\Psi[x(n)]}{1 - \dfrac{[x(n+1) + x(n-1)]^2}{4x^2(n)}}}. \tag{164}$$

This algorithm has the advantage that it contains no signal differences; instead, it averages. Hence, it may be more robust. However, it does not generalize well to the time-varying case. Specifically, let $x(n) = a(n) \cos[\phi(n)]$ be an AM-FM signal. Then we could use the formulas (163) and (164) to approximately estimate the time-varying amplitude envelope $|a(n)|$ and instantaneous frequency $\Omega_i(n)$ provided that the following conditions were true; i) $a(n)$ is bandlimited with bandwidth $\Omega_a << \Omega_c$, ii) $\Omega_i(n)$ is bandlimited with bandwidth $\Omega_f << 1$ and has deviation $\Omega_m << \Omega_c$, and iii) $\sin(\Omega_a) << \cos(\Omega_c)$. Although conditions i) and ii) are quite realistic, condition iii) presents a serious problem: It is true only if Ω_c is close to zero or to π, i.e., for very small or very high carrier frequencies. For intermediate frequencies Ω_c around $\pi/2$, condition iii) cannot be generally true.

3) Hilbert Transform: Given a real AM–FM signal $x(t) = a(t)\cos[\phi(t)]$, an alternative approach to estimate its envelope $|a(t)|$ and instantaneous frequency $\omega_i(t) = \dot{\phi}(t)$ is to use the Hilbert transform of $x(t)$. Specifically, if $X(\omega)$ is the Fourier transform of $x(t)$, its Hilbert transform is the signal $\hat{x}(t)$ with Fourier transform $\hat{X}(\omega) = -j\,\mathrm{sgn}(\omega)X(\omega)$. The related complex-valued analytic signal is

$$\text{analytic signal: } x(t) + j\hat{x}(t) = r(t)e^{j\theta(t)} \tag{165}$$

with $r(t) \geq 0$. Thus, the Hilbert transform can provide an envelope $r(t) = \sqrt{x^2(t) + \hat{x}^2(t)}$ and an instantaneous frequency $\dot{\theta}(t)$, where $\theta(t) = \arctan[\hat{x}(t)/x(t)]$. In general, $r(t)$ and $\dot{\theta}(t)$ will be different from their counterparts $|a(t)|$ and $\omega_i(t)$ imposed by the AM–FM model. However, conditions exist [23] that guarantee that the error between the Hilbert transform $\hat{x}(t)$ and the quadrature signal $a(t)\sin[\phi(t)]$ is small. Similar arguments can be developed to compare $r(t)$ and $\dot{\theta}(t)$ with $|a(t)|$ and $\omega_i(t)$.

For *discrete-time* signals a Hilbert transform can be approximately implemented using either the discrete Fourier transform (i.e., FFT) or an FIR filter, as explained in [24]. In a comparative work [25], experiments on N-sample synthetic AM–FM signals indicate that when the ratios of carrier Ω_c versus the information signals' bandwidths Ω_a and Ω_f are in the order of 10, as is the case in speech applications, then the Hilbert transform (implemented via an FIR filter) can give a smaller error but at a computational complexity $O(N^2)$ which is higher than the very low $O(N)$ complexity of the DESA. (Also the complexity of the Hilbert transform implemented using FFT's is $O(N \log_2 N)$.) Decreasing the complexity of the Hilbert transform to an $O(N)$ by using a shorter impulse response makes its error larger than that of the DESA. In addition, if the ratios of Ω_c versus Ω_a and Ω_f are in the order of 100 or higher, as in communications applications, then the DESA yields a smaller error than the more complex Hilbert transform approach. Experiments on real speech signals indicate that both approaches yield similar estimates. In very few cases, the Hilbert transform yielded somewhat smoother amplitude and frequency signals, but a very short FIR smoothing of the energy signals makes the DESA achieve similar smoothness and still maintain lower complexity [25]. Another advantage of the DESA is that (for each output sample) it uses an extremely short window (of 5 samples) which allows it to instantaneously adapt during speech transitions between phonemes, whereas the Hilbert transform often needs a window whose length is of the same order as the length of the speech analysis frame.

C. General Conclusions

In this paper we have developed a theory for estimating the time-varying amplitude envelope and instantaneous frequency of arbitrary AM–FM signals using nonlinear energy-tracking operators. The only constraint imposed is that the amplitude and frequency signals do not vary too fast or too greatly with time compared with the carrier. The continuous-time theoretical analysis is accompanied with the development of efficient energy separation algorithms for discrete-time signals. We have developed three DESA's and numerically compared their performance on synthetic AM–FM signals. We have found that the DESA's work quite effectively for estimating the amplitude envelope and instantaneous frequency under fairly broad ranges of amounts of amplitude and/or frequency modulation. In addition, their computationally complexity is very small, i.e., linear in the number of signal samples with a small proportionality constant.

We have also applied the DESA's to the analysis of speech resonances, modeled using AM–FM signals. The DESA's uncovered interesting AM and FM structure in signals from speech resonances within a single pitch period. Speech formants with time-varying nonexponential amplitude envelopes and oscillating instantaneous values have been frequently observed. These results support a general model for a short-time speech segment as a superposition of AM–FM resonances, as proposed in [18], [19]. However, our experimental results on speech are only a beginning. Much more work remains to be done in refining this AM–FM model for speech and the estimation of its parameters. In the meantime, we strongly believe that the energy operators and ESA's, due to their simplicity and instantaneous-adapting nature, offer valuable signal processing tools for detecting interesting modulation patterns in speech and other time-varying signals.

APPENDIX

The error analysis in this appendix uses the same type of approximations as in [18].

Let us assume that $a(t)$ and $\dot{\phi}(t) = \omega_i(t)$ are bandlimited with bandwidths $\omega_a \ll \omega_c$ and $\omega_f \ll \omega_c$, respectively. Also assume $\omega_m \ll \omega_c$.

By physical order of magnitude $\mathcal{O}(c)$ for a constant c we shall mean the power of 10 closest to c. Due to the oscillatory nature of the signals analyzed herein, we define the *order of magnitude of a signal* $x(t)$ to be the order of magnitude of its maximum absolute value; i.e.,

$$\mathcal{O}(x(t)) \triangleq \mathcal{O}(x_{\max}). \tag{166}$$

For example,

$$\mathcal{O}(\omega_i) = \mathcal{O}(\omega_c + \omega_m). \tag{167}$$

Since we will compare dominant and error terms with very different orders of magnitude, we use the following approximate laws for $\mathcal{O}(\,)$: for $A, B > 0$

$$\mathcal{O}(AB) \approx \mathcal{O}(A)\mathcal{O}(B)$$

$$\mathcal{O}(A + B) \approx \max[\mathcal{O}(A), \mathcal{O}(B)]. \tag{168}$$

In this paper, we assume that

$$\mathcal{O}(x_{\max}) \approx \mathcal{O}(\mu_x) \tag{169}$$

for bandlimited signal $x(t)$. This implies that for the amplitude a and the frequency modulating signal q we have $\mathcal{O}(a) \approx \mathcal{O}(\mu_a)$ and $\mathcal{O}(q) = \mathcal{O}(1)$.

To find the order of the error terms in $\Psi(\dot{x}) - a^2\omega_i^4$, note that combining (20)–(22) yields

$$\Psi(\dot{x}) = a^2\omega_i^4 + E \tag{170}$$

where E is the total approximation error equal to

$$E = (\dot{y}_1)^2 - y_1\ddot{y}_1 - 2\dot{y}_1\dot{y}_2 + y_1\ddot{y}_2 + \ddot{y}_1 y_2 - 0.5a^2\dot{\phi}^2\ddot{\phi}\sin(2\phi) + [a^2\Psi(\dot{\phi}) + \dot{\phi}^2\Psi(a)] \cdot \sin^2(\phi) \tag{171}$$

and, by (20),

$$\dot{y}_1 = \dot{a}\cos\phi - a\dot{\phi}\sin\phi \tag{172}$$

$$\ddot{y}_1 = (\ddot{a} - a\dot{\phi}^2)\cos\phi - (2\dot{a}\dot{\phi} + a\ddot{\phi})\sin\phi \tag{173}$$

$$\dot{y}_2 = (\dot{a}\dot{\phi} + a\ddot{\phi})\sin\phi + a\dot{\phi}^2\cos\phi \tag{174}$$

$$\ddot{y}_2 = (\ddot{a}\dot{\phi} + 2\dot{a}\ddot{\phi} + a\dddot{\phi} - a\dot{\phi}^3)\sin\phi + (2\dot{a}\dot{\phi}^2 + 3a\dot{\phi}\ddot{\phi})\cos\phi. \tag{175}$$

From Lemma 1 and (33) it follows that

$$\mathcal{O}(\dot{a}) \approx \mathcal{O}(a\omega_a) \tag{176}$$

$$\mathcal{O}(\ddot{a}) \approx \mathcal{O}(a\omega_a^2) \tag{177}$$

$$\mathcal{O}(\ddot{\phi}) \approx \mathcal{O}(\omega_m\omega_f) \tag{178}$$

$$\mathcal{O}(\dot{y}_1) \approx \mathcal{O}(a\omega_a\omega_i) \tag{179}$$

$$\mathcal{O}(\dot{y}_2) \approx \mathcal{O}(a\omega_i^2) \tag{180}$$

$$\mathcal{O}(\ddot{y}_1) \approx \mathcal{O}(a\omega_a\omega_i^2) \tag{181}$$

$$\mathcal{O}(\ddot{y}_2) \approx \mathcal{O}(a\omega_i^3). \tag{182}$$

Hence, the orders of the terms in E are

$$\mathcal{O}(\dot{y}_1^2) \approx \mathcal{O}(a^2\omega_a^2\omega_i^2) \tag{183}$$

$$\mathcal{O}(y_1\ddot{y}_1) \approx \mathcal{O}(a^2\omega_a^2\omega_i^2) \tag{184}$$

$$\mathcal{O}(\dot{y}_1\dot{y}_2) \approx \mathcal{O}(a^2\omega_a\omega_i^3) \tag{185}$$

$$\mathcal{O}(y_1\ddot{y}_2) \approx \mathcal{O}(a^2\omega_a\omega_i^3) \tag{186}$$

$$\mathcal{O}(\ddot{y}_1 y_2) \approx \mathcal{O}(a^2\omega_a\omega_i^3) \tag{187}$$

$$\mathcal{O}(a^2\dot{\phi}^2\ddot{\phi}) \approx \mathcal{O}(a^2\omega_i^2\omega_m\omega_f) \tag{188}$$

$$\mathcal{O}[a^2\Psi(\dot{\phi})] \approx \mathcal{O}(a^2\omega_f^2\omega_i^2) \tag{189}$$

$$\mathcal{O}[\dot{\phi}^2\Psi(a)] \approx \mathcal{O}(a^2\omega_a^2\omega_i^2). \tag{190}$$

Thus, the order of the total error is

$$\mathcal{O}(E) \approx \mathcal{O}(a^2)\max[\mathcal{O}(\omega_a\omega_i^3), \mathcal{O}(\omega_m\omega_f\omega_i^2), \mathcal{O}(\omega_f^2\omega_i^2)]. \tag{191}$$

Since the desired energy term is $a^2\omega_i^4$ and ω_a, ω_m, $\omega_f \ll \omega_c$, it follows that $\mathcal{O}(E) \ll \mathcal{O}(a^2\omega_i^4)$. Thus, the dominant term in (170) is $a^2\omega_i^4$, which yields the final approximation (40).

ACKNOWLEDGMENTS

The authors would like to thank Helen Hanson and Alexandros Potamianos at Harvard University for many insightful discussions and experiments on applying the DESA's to speech analysis.

REFERENCES

[1] E. H. Armstrong, "A method of reducing disturbances in radio signaling by a system of frequency modulation," *Proc. IRE*, vol. 24, pp. 689–740, May 1936.

[2] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[3] J. R. Carson, "Notes on the theory of modulation," *Proc. IRE*, vol. 10, pp. 57–64, 1922.

[4] L. Deng, P. Kenny, M. Lennig, and P. Mermelstein, "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 265–271, Feb. 1992.

[5] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960, 1970.

[6] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. Berlin: Springer-Verlag: 1965, 1972.

[7] D. Gabor, "Theory of communication," *IEE J.*, London, vol. 93, pp. 429–457, 1946.

[8] H. M. Hanson, P. Maragos, and A. Potamianos, "A system for finding speech formants and modulations via energy separation," Tech. Rep. 92-6, Harvard Robotics Lab, Harvard Univ., June 1992.

[9] G. C. Hegerl and H. Höge, "Numerical simulation of the glottal flow by a model based on the compressible Navier–Stokes equations," in *Proc. IEEE ICASSP-91*, Toronto, Canada, May 1991, pp. 477–480.

[10] H. Iijima, N. Miki, and N. Nagai, "Fundamental consideration of finite element method for the simulation of the vibration of vocal cords," in *Proc. IEEE ICASSP-89*, Glasgow, Scotland, May 1989, pp. 246–249.

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

[12] ——, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE ICASSP-90*, Albuquerque, NM, Apr. 1990, pp. 381–384.

[13] ——, "On Teager's energy algorithm and its generalization to continuous signals," in *Proc. 4th IEEE Digital Signal Processing Workshop*, Mohonk, New Paltz, NY, Sept. 1990.

[14] G. E. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 709–729, Aug. 1986.

[15] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[16] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin: Springer-Verlag, 1976.

[17] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Detecting nonlinearities in speech using an energy operator," in *Proc. 4th IEEE Digital Signal Processing Workshop*, Mohonk, New Paltz, NY, Sept. 1990.

[18] ——, "Speech nonlinearities, modulations, and energy operators," in *Proc. IEEE ICASSP-91*, Toronto, Canada, May 1991, pp. 421–424.

[19] P. Maragus, J. F. Kauer, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Processing*, vol. 41, pp. 1532–1550, Apr. 1993.

[20] J. S. Marques and L. B. Almeida, "Frequency-varying sinusoidal modeling of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 763–765, 1989.

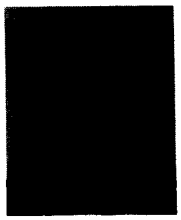
[21] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.

[22] R. S. McGowan, "An aeroacoustics approach to phonation," *J. Acoust. Soc. Amer.*, vol. 83, no. 2, pp. 696–704, Feb. 1988.

[23] A. H. Nuttall, "Complex envelope properties, interpretation, filtering, and evaluation," NUSC Tech. Rep. 8827, New London, CT, Feb. 1991.

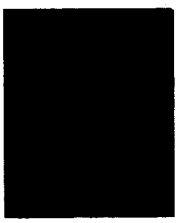
[24] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[25] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," Tech. Rep., 92-8, Harvard Robotics Lab, Harvard Univ., July 1992.
[26] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[27] M. Schwartz, *Information Transmission, Modulation and Noise*. New York: McGraw-Hill, 1980.
[28] J. Shekel, "Instantaneous frequency," *Proc. IRE*, vol. 41, p. 548, 1953.
[29] M. M. Sondhi and J. Schroeter, "Articulatory speech analysis/synthesis," *IEEE ICASSP-92*, Tutorial, San Francisco, CA, Mar. 1992.
[30] H. M. Teager, "Screening for vocal tract neoplasms," proposal submitted, 1976.
[31] ——, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 5, pp. 599-601, Oct. 1980.
[32] ——, "Physiology of speech phoneme production," 3rd Tech. Rep., Contract MDA 904-81-C-0413, 1981.
[33] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract," in R. G. Daniloff, Ed., *Speech Sciences: Recent Advances*. San Diego, CA: College-Hill Press, pp. 73-109, 1983.
[34] ——, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, William J. Hardcastle and Alain Marchal, Eds. (NATO Advanced Study Institute Series D, vol. 55, Bonas, France, July 17-29, 1989). Boston, MA: Kluwer, pp. 241-261, 1990.
[35] T. J. Thomas, "A finite element model of fluid flow in the vocal tract," *Comput. Speech Language*, vol. 1, pp. 131-151, 1986.
[36] D. J. Tritton, *Physical Fluid Dynamics*, 2nd ed. New York: Oxford Univ. Press, 1988.
[37] B. van der Pol, "Frequency modulation," *Proc. IRE*, vol. 18, pp. 1194-1205, July 1930.
[38] ——, "The fundamental principles of frequency modulation," *IEE J.*, London, vol. 93, pp. 153-158, 1946.
[39] 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.



Petros Maragos (S'81-M'85-SM'91) was born in Kalymnos, Greece, on November 4, 1957. He received the Diploma degree in electrical engineering from the National Technical University of Athens, Greece, in 1980, and the M.S.E.E. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, in 1982 and 1985 respectively.

In 1985 he joined the faculty of the Division of Applied Sciences at Harvard University, Cambridge, MA, where he worked as Assistant, from 1985 to 1989, and Associate Professor of Electrical Engineering from 1989 to 1993. During the fall of 1992 he was a Visiting Professor at the National Technical University of Athens. In June 1993, he joined the faculty of Georgia Tech, Atlanta, GA. His general research and teaching activities have been in the areas of signal processing, image processing and computer vision, computer speech processing and recognition, and applied mathematics. His current research focuses on nonlinear systems for speech and image processing, morphological signal analysis, fractals, pattern recognition, and communications. Dr. Maragos has been involved in many professional society activities, including: Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1989 to 1992; Editorial Board Member for the *Journal of Visual Communication and Image Representation* from 1990 to 1992; General Chairman for the SPIE Conference on Visual Communications and Image Processing, Boston, November 1992; Member of the IEEE Digital Signal Processing Technical Committee from 1992 to present. He received a Sigma Xi thesis research award in 1983 for his work on image coding. In 1987, he received a National Science Foundation Presidential Young Investigator Award for his work in signal and image processing. He is also the recipient of the IEEE Acoustics, Speech, and Signal Processing Society's 1988 Paper Award for the paper "*Morphological Filters-Parts I and II*" published in the Society's Transactions.



James F. Kaiser (S'50-A'52-SM'70-F'73) was born in Piqua, OH, in 1929. He received the E.E. degree from the University of Cincinnati, Cincinnati, OH, in 1952 and the S.M. and Sc.D. degrees in 1954 and 1959 respectively, from M.I.T., Cambridge, MA, all in electrical engineering.

Currently he is a Visiting Professor in the Department of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ. He was formerly a Distinguished Member of Technical Staff in the Speech and Image Processing Research Division of Bell Communications Research, Inc., which he joined in 1984 at its formation. Prior to that he was a Distinguished Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, for 25 years where he worked in the areas of speech processing, system simulation, digital signal processing, computer graphics, and computer-aided design. He also serves as a consultant to industry. He is the author of more than 60 research papers and the coauthor and editor of seven books in the signal processing and automatic control areas. He has served in a number of positions in both the Signal Processing Society and Circuits and Systems Society including Administrative Committee member, Secretary-Treasurer, Digital Signal Processing Technical Committee member, and Awards Board Chair and member. He has also served on the IEEE Press Board and as an ECPD Visitor. For six years he was Secretary of the COSINE Committee of the Commission on Education of the National Academy of Engineering.

Dr. Kaiser is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He has also received the Technical Achievement Award of the Signal Processing Society in 1978, its Meritorious Service Award in 1979, its Society Award in 1982, and the IEEE Centenial Medal in 1984. In 1970 he was presented with the Distinguished Engineering Alumnus Award by the University of Cincinnati, College of Engineering and in 1980 the Eta Kappa Nu Award of Merit also by the University of Cincinnati. He is a Registered Engineer in Massachusetts, a member of ASA, AAAS, EURASIP, and SIAM.

Thomas F. Quatieri (S'73-M'79-SM'87) was born in Somerville, MA on January 31, 1952. He received the B.S. degree (summa cum laude) from Tufts University, Medford MA, in 1973, and the S.M., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology (M.I.T.), Cambridge, MA in 1975, 1977, and 1979, respectively.

In 1980, he joined the Sensor Processing Technology Group of M.I.T., Lincoln Laboratory, Lexington, MA where he worked on problems in multidimensional digital signal processing and image processing. Since 1983 he has been a member of the Speech Systems Technology Group at Lincoln Laboratory where he has been involved in digital signal processing for speech coding and enhancement, underwater sound enhancement, and data communications. He has contributed many publications to journals and conference proceedings, written several patents, and co-authored chapters in two edited books: *Advanced Topics in Signal Processing* (Prentice-Hall, 1987) and *Advances in Speech Signal Processing* (Marcel-Dekker, 1991). He is also the Lecturer for the MIT graduate course *Digital Speech Processing*.

Dr. Quatieri is the recipient of the 1982 Paper Award of the IEEE Acoustics, Speech, Signal Processing Society for the paper, "Implementation of 2-D Digital Filters by Iterative Methods." In 1990, he received the IEEE Signal Processing Society's Senior Award for the paper, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," published in the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. He is a member of the IEEE Digital Signal Processing Technical Committee, and from 1983 to 1992 has served on the steering committee for the bi-annual Digital Signal Processing Workshop, and is currently Associate Editor for the *IEEE Transactions on Signal Processing* in the area of nonlinear systems. He is also a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.