On Decomposing Speech into Modulated Components

Ashwin Rao, *Member, IEEE,* and Ramdas Kumaresan, *Fellow, IEEE*

***Abstract*—We model a segment of filtered speech signal as a product of elementary signals as opposed to a sum of sinusoidal signals. Using this model, one can better appreciate the basic relationships between envelopes and phases or instantaneous frequencies (IF's) of signals. These relationships reveal some interesting properties of the signal's modulations. For instance, if the contribution due to a signal's envelope, specifically the Hilbert transform of its log-envelope, is removed from the signal's phase, then the resulting signal's IF is strictly positive. In addition, filtered speech signal having a bandwidth of B Hz can be essentially represented by log-envelope and IF that have the same B Hz bandwidths. In this paper, we extend the above ideas to decompose speech into modulated components. Specifically, a bank of data-adaptive filters (in a cross-coupled configuration) are used to decompose speech into its components; each adaptive filter is a simple single-resonance bandpass filter (whose center-frequency or pole-location closely follows the desired formant frequency) supplemented by an adaptive all-zero filter (whose zero-locations sufficiently reduce unwanted leakage from neighboring formants). The filtered components are then represented by their respective log-envelopes and positive IF's; these small number of modulations closely approximate the speech signal.**

***Index Terms*—Auditory model, formant, front-end, modulations, product representation, speech features.**

I. Introduction

CURRENTLY popular speech processing techniques are almost exclusively based on spectral analysis in the form of linear prediction [1], [2], cepstral analysis [3], and Mel-cepstrum [4]. In one way or another, they invoke the source-filter model of speech generation [5]. Using these procedures, spectral templates or feature vectors are computed and used in applications like machine recognition/verification, synthesis, and coding of speech. Unfortunately, when speech is accompanied with noise, reverberation, and other degradations, perturbations at one frequency affect the entire template rendering the extracted features vulnerable [6]. Other procedures, such as spectrogram or fixed filter bank processing, in which the signal energy is computed as a function of frequency, have similar drawbacks; significant energy from a nearby filter (of the filter-bank) biases the estimate of energy in a desired filter. Also, these techniques practically discard all phase information in a signal's spectrum. In order to improve the performance of many systems associated with speech applications, improved speech processing methods are necessary [7]. In this paper, we propose a new way of representing and analyzing speech. It deviates from the traditional spectral representation based techniques, by decomposing speech into handful of smooth modulated components.

We propose to decompose a given speech signal into modulations, namely the signal's envelopes and IF's. An analytic signal approach is adopted because it permits an unambiguous characterization of a real signal in terms of its envelope and IF. Recall that if $s_r(t)$ is a real signal, then the corresponding analytic signal is $s(t) = s_r(t) + j\hat{s}_r(t)$, where $\hat{s}_r(t)$ is the Hilbert transform of $s_r(t)$. The envelope of $s_r(t)$ is then defined as the magnitude of $s(t)$ (denoted as $|s(t)|$) and its IF is defined as the first derivative of $s(t)$'s phase function scaled by $1/2\pi$. Characterizing signals by envelopes and IF's are also commonly referred to as AM–FM modeling of signals.

Related speech processing methods that attempt to extract modulation information from a speech signal may be traced back to early works such as in Dudley [8] and Cherry and Phillips [9]. Flanagan [10] used analytic signals to characterize bandpass speech in terms of envelopes and IF's. Extraction of "formant modulations" has also been addressed in the context of vocal-tract modeling. For instance, in [11] Atal attempted to decompose a speech wave into signals representing vocal-tract resonance waveforms. Other proposed variants include the works published in [12] and [13]. One of the drawbacks in these earlier procedures is that leakage from neighboring formants affects the estimate of a desired formant's location, resulting in significant variability in the extracted formant modulations. To address this problem, Jackson and Bertrand have proposed [14] an adaptive inverse filtering approach for formant analysis. The algorithm extends the well-known least-mean-squared (LMS) algorithm in a cascade form to estimate inverse filters on a sample-by-sample basis and subsequently to track individual formants. However, the procedures mentioned above merely address estimation of the slowly varying formant frequencies and their respective bandwidths. Generally, the variations (or details) in amplitude and phase/frequency (that characterize the many harmonics buried under a formant) are discarded by smoothing the corresponding quantities. The reason for this may be attributed to the difficulties associated with understanding and estimating the IF (and log-envelope) of an arbitrary signal like speech.

Electrical engineers are most familiar with IF in the context of frequency modulated signals as in an FM radio. But what about the IF of a signal consisting of several tones as in vowel

Manuscript received October 22, 1997; revised June 17, 1999. This work was supported by the National Science Foundation under Grant CCR-9804050. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerard Chollet.

A. Rao is with Dragon Systems, Inc., Newton, MA 02460 USA (e-mail: ashwin_rao@dragonsys.com).

R. Kumaresan is with the Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881 USA (e-mail: kumar@ele.uri.edu).

Publisher Item Identifier S 1063-6676(00)03452-0.

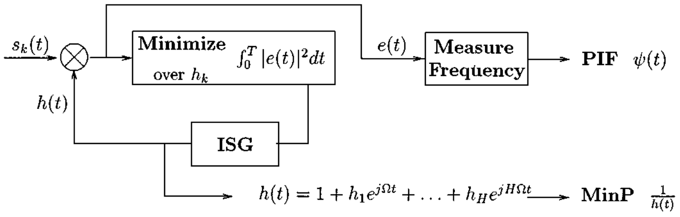


Fig. 1. LPSD algorithm: $1/h(t)$ [where $h(t)$ is the output of the inverse signal generator, ISG] corresponds to the MinP part of the signal $s_k(t)$. $\psi(t)$ corresponds to the IF of the AllP part of the signal $s_k(t)$.

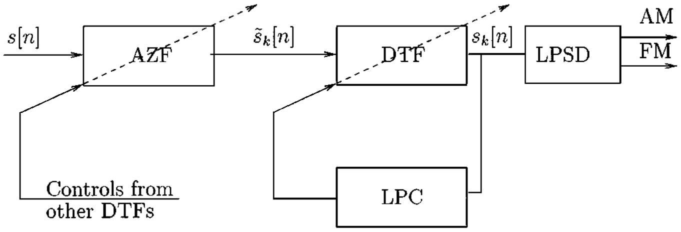


Fig. 2. The DTF is shown as part of the block-diagram for the kth channel of the adaptive filterbank. The DTF's center-frequency is estimated using a single-pole LP analysis (discussed later).

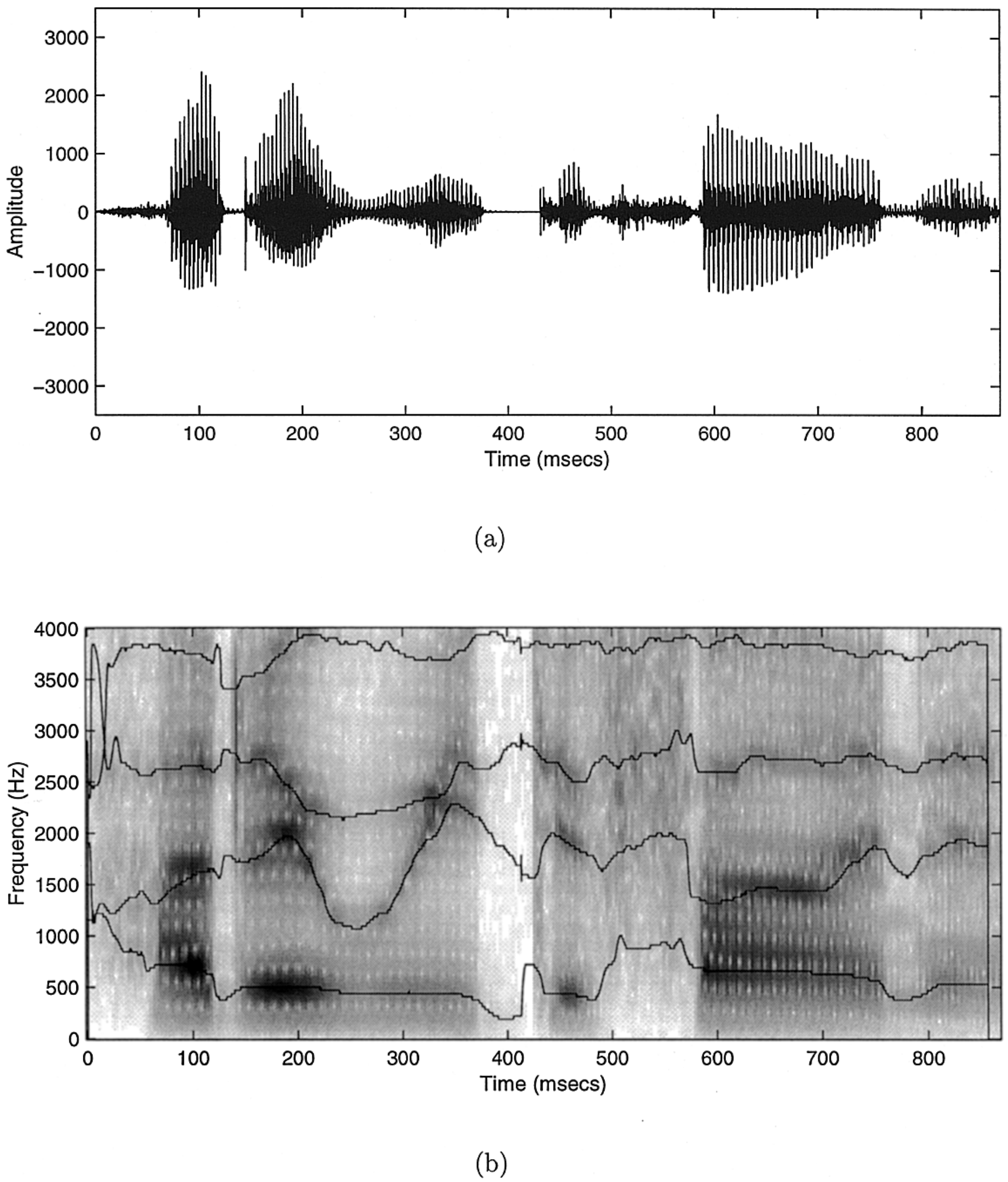
sounds? For an arbitrary signal, the IF is typically an erratic function whose range may extend from negative to positive infinity [15]. The general impression among researchers is that the IF function is unusable unless it is sufficiently smoothed [10]. Nearly 30 years ago, Voelcker [16] proposed a methodical way that leads to understanding the IF (and log-envelope) of signals. He implies in his arguments that describing signals using sum of sinusoidal signals as in Fourier analysis is unsuitable for time localized description of signals. Instead, he proposed that complex-valued signals (and hence analytic signals) may be modeled as a ratio of polynomials in the complex variable t (time), just like a given system or frequency response may be modeled as a ratio of polynomials in the s-domain or the z-domain. That is, a bandlimited periodic analytic signal $s(t)$, with period T s ($\Omega = 2\pi/T$ denotes its fundamental angular frequency) may be described by the following formula for a sufficiently large P and Q:

$$s(t) = a_0 e^{j\omega_t t} \prod_{i=1}^{P} (1 - p_i e^{j\Omega t}) \prod_{i=1}^{Q} (1 - q_i e^{j\Omega t}) \qquad (1)$$

where $e^{j\omega_t t}$ corresponds to a frequency translation ω_t. The products represent the complex envelope of the signal. p_i and q_i correspond to zeros on or inside and outside the unit circle in the complex time plane, respectively. More generally, band unlimited signals may have both poles and zeros. Voelcker called this way of modeling signals as product representation of signals. Analogous to the unit-circle in the (discrete-time) z-plane [3] which corresponds to a frequency range of 0–2π, the unit-circle in the complex time plane corresponds to a time duration of 0–T s. Using this product representation model, it is easy to understand the relationship between phase and envelope of signals. For example, if a periodic signal is such that a zero of the signal is close to the unit circle, then significant phase changes will occur in the temporal neighborhood of this zero which will be reflected in the IF values. Specifically, a zero or a pole close to the unit-circle will result in a large spike in the IF. In fact, if a zero happens to fall on the circle, the envelope goes to zero (at a time instant determined by the zero's location) and the IF at that time instant is undefined (á la group delay of systems). Thus, using the above representation a number of familiar results in systems theory can be used to understand relationships between IF and log-envelope of signals.

In recent works [17]–[20], we have extended Voelcker's idea. Once we realize that signals may be represented over a short interval of T s by a ratio of polynomials with complex coefficients, then many ideas that have been developed in systems literature can be applied to this so-called product representation of signals. In [17]–[19], we devised a method to approximate a signal's envelope using a minimum phase signal model. This approach is analogous to the standard linear prediction method well known in spectral analysis as "autocorrelation method" [21]. This method leads to a general technique for decomposing an analytic signal into a minimum phase (MinP) signal and an all-phase (AllP) signal. A MinP signal will have all its zeros inside the unit-circle and is completely characterized by its envelope alone and the AllP signal has a positive definite IF. The significance of our result is that the MinP-AllP decomposition is achieved without actually finding the zeros of the signal; the result is a simple procedure that acts as a unique positive FM–AM adaptive demodulator. In this paper, we exploit this decomposition and other related properties of AM–FM of multicomponent sinusoidal signals to decompose speech signals into modulated components.

The paper is organized as follows. In Section II, we describe the signal model for a suitably filtered speech signal. Brief insights into log-envelope and IF of our modeled signal are also provided in this same section; these discussions may be viewed as an extension of our recently published works [17]–[20] for modeling speech. In Section III, the linear prediction in spectral domain (LPSD) algorithm that we have proposed in [19], [17], and [18] is reviewed; in Section IV the LPSD algorithm is used to decompose filtered speech into envelope and positive IF. In Section IV, we first motivate the need for a time-varying filterbank for processing speech. Our arguments are different from earlier published works, since they are based on the signal model described in Section II. Following this, we provide an automatic algorithm for designing such an adaptive filterbank. Specifically, the filter bank is composed of several single pole filters whose pole angles are adaptively changed based on center frequencies' estimates obtained using a first-order linear prediction analysis. They are different from some of our earlier works that used similar strategies (albeit different IF-estimation techniques) for tracking multiple narrowband (slowly varying tones/harmonics) components. Another difference between the proposed algorithm and previously published works is that our procedure relies on adaptively suppressing regions of strong spectral contents (neighboring formants) while filtering (or tracking) a desired spectral-band (or desired formant). The results of simulations using real speech waveforms are provided in Section V, followed by discussions and conclusion.



(a)

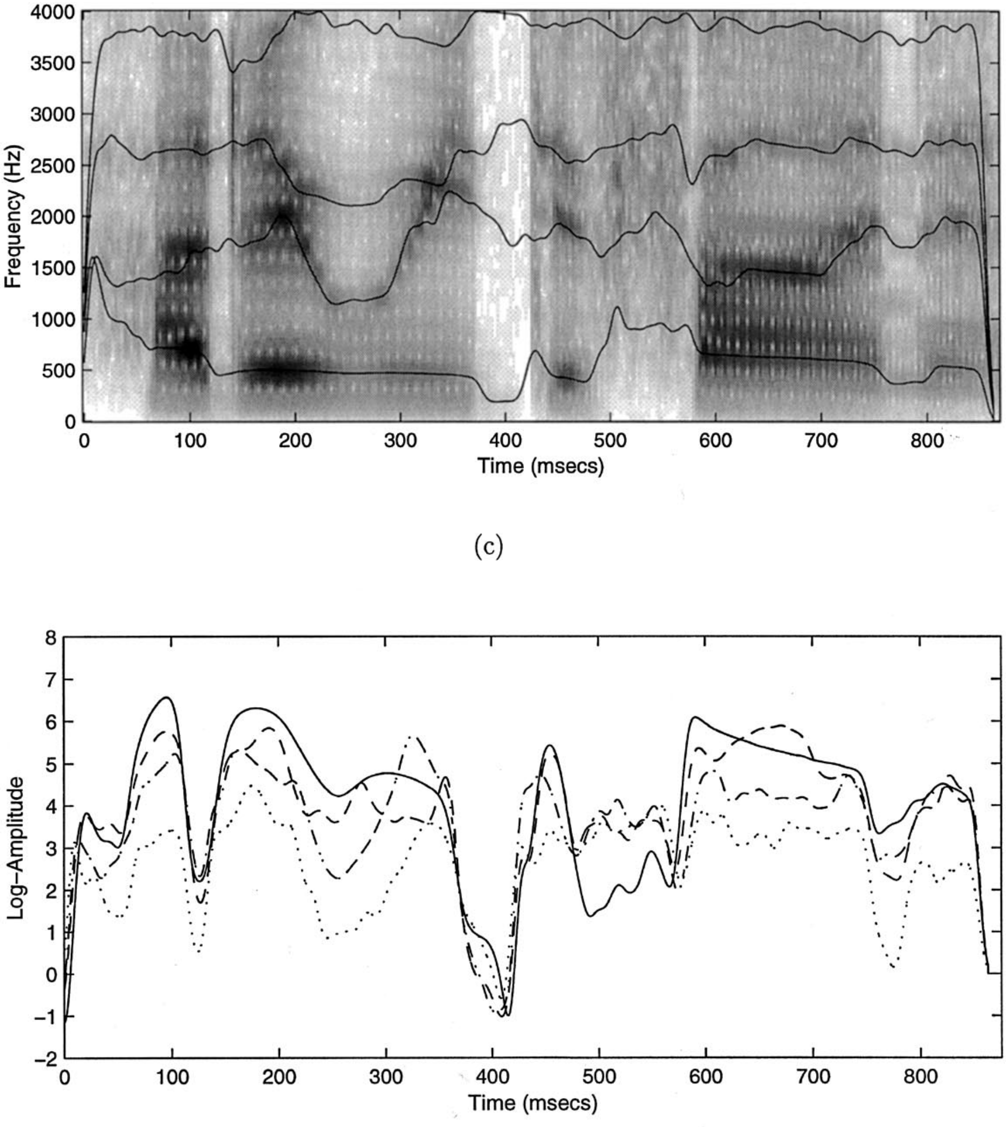
(b)

Fig. 3. The speech signal ("How do we define it") is shown in (a) while (b) corresponds to its spectrogram. Also, in (b) the DTF's center frequencies estimated using a single-pole LP analysis are overlaid. Observe that they closely follow the formant frequencies during the voiced-part of the signal.

II. Signal Model for Filtered Speech

There is no unique way of decomposing an arbitrary signal like speech and representing its components using envelope and IF (AM–FM) information, unless some constraints are imposed. At one extreme lies the harmonic decomposition scheme wherein every individual harmonic component in speech is modeled as an AM–FM component [22]–[25]. It is typically achieved by computing the envelope and estimating the IF at the outputs of several narrow bandpass filters. Accurately tracking the envelope and IF of individual components (harmonics in voiced speech) has also been attempted [26], [27]. This effort is complicated by the fact that many harmonic components in a speech signal (specifically voiced-speech) are transient in nature, i.e., the harmonics are subject to a "birth-death" like phenomenon [25] as they move in and out of a formant region. The other extreme is to have no filters at all and hence represent the signal in its entirety using a single AM–FM representation. The former case will typically result in a large number of smooth modulations; although some early attempts to track harmonic partials in voiced speech [27], [28] have demonstrated that for brief time instances the harmonics exhibit significant departure from strict harmonicity. The latter will yield a single but extremely wild AM and FM [19]. It appears that a reasonable strategy to parsimoniously represent a signal (speech in our case) is to filter the signal through a bank of filters such that a compromise is achieved between the smoothness of envelope and frequency modulations and the number of filters used in the decomposition.

Constant-Q filterbanks (wherein filters centered at higher frequencies are designed to have larger bandwidths) have been used in some speech applications [4]. These were inspired by the phenomenon of critical bands observed in the



(c)

(d)

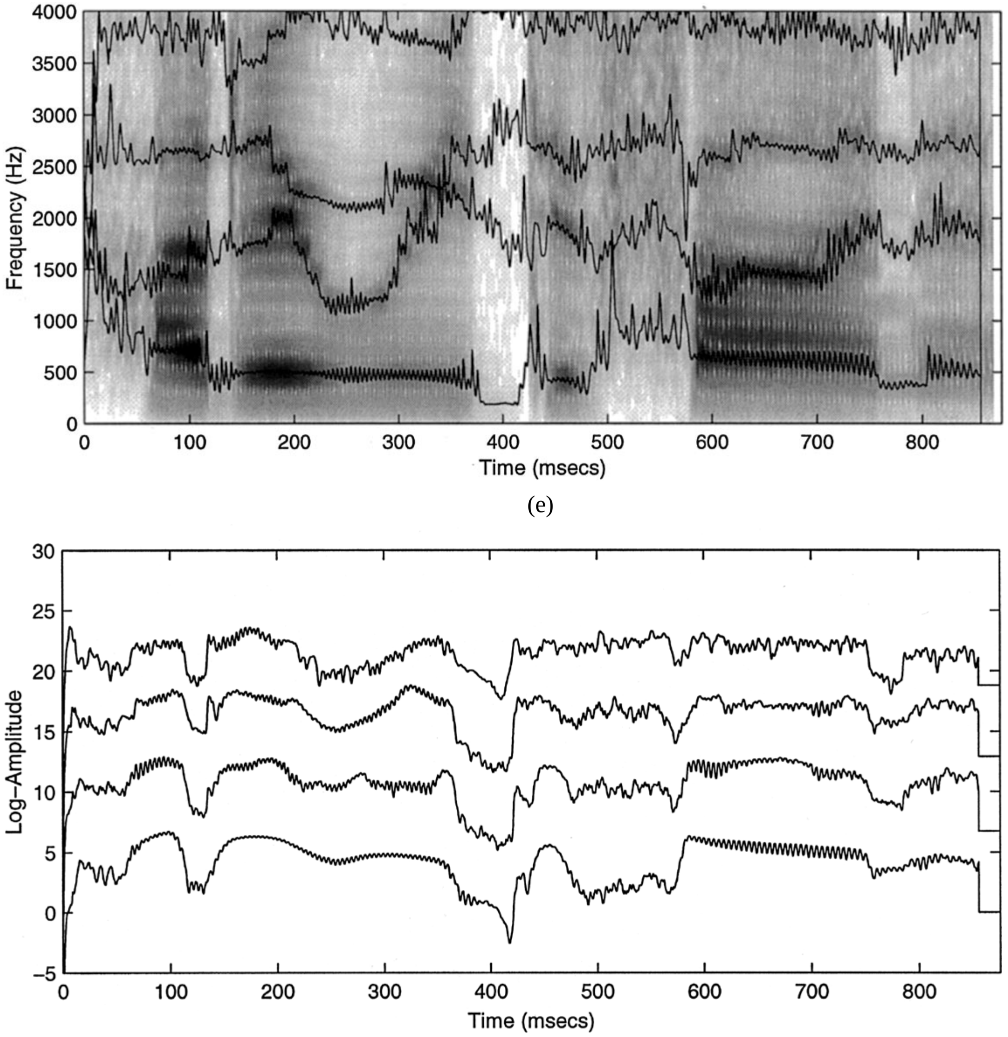
Fig. 3. (*Continued.*) (c) Slowly varying carrier frequencies and (d) the smooth log-envelopes of the components forming the DTF's outputs. The carrier frequencies are shown along with the spectrogram. The smooth log-envelopes are plotted using solid, dashed, dotted, and dashed–dotted lines corresponding to log-envelopes of components at the output of DTF's 1–4, respectively.

auditory system [29]. Quatieri *et al.* [30] have used other filter shapes such as gammatone, Gaussian, and piecewise linear, which are also motivated by filtering in the early stages of auditory processing. However, such filterbanks do not guarantee that the filtered outputs will have smooth envelope and IF variations. The key appears to be in making the filters adapt such that the envelope and IF of each filter output is smooth. Thus, we focus our efforts in trying to achieve such an adaptive filter. But before we proceed in that direction we set up the notation for a signal model that describes the filtered speech signal in terms of its envelope and IF.

We shall assume that $s(t)$ is the analytic signal formed from the real-valued speech signal. Then, $s(t)$ is filtered through a bank of complex filters. For the time-being, let us further assume that the filters are linear and time-invariant. Let $s_k(t)$ be the output of the kth filter having a finite bandwidth of B Hz. We then set $T = 1/B$ s to be the processing interval for $s_k(t)$. As in traditional short-time spectral analysis, the filtered signal $s_k(t)$ is assumed to be periodically replicated. The period is T s. Let $\Omega = 2\pi/T$ denote the fundamental angular frequency. Since the kth filter (centered around a region of significant spectral energy) is typically a bandpass filter, the spectrum of $s_k(t)$ will be concentrated around the center frequency of this filter. Then for a sufficiently large M, over a time interval of T s, $s_k(t)$ can be modeled as

$$s_k(t) = e^{j\omega_t t} \sum_{k=0}^{M} a_k e^{jk\Omega t} \tag{2}$$

where a_k's are the complex amplitudes of the sinusoids. $a_0 \neq 0$ and $a_M \neq 0$. $e^{j\omega_t t}$ represents the frequency translation. The



(e)

(f)

Fig. 3. (*Continued.*) In (e) and (f), we display the modulations: the PIF's and the log-envelopes are shown in (e) and (f), respectively. The PIF's are plotted along with the spectrogram. The log-envelopes are displayed on top of each other, starting with the first DTF's modulation (at the bottom).

Mth degree polynomial in $e^{j\Omega t}$ describes the complex envelope of the signal $s_k(t)$. We may factor this polynomial into its M factors and rewrite $s_k(t)$ as

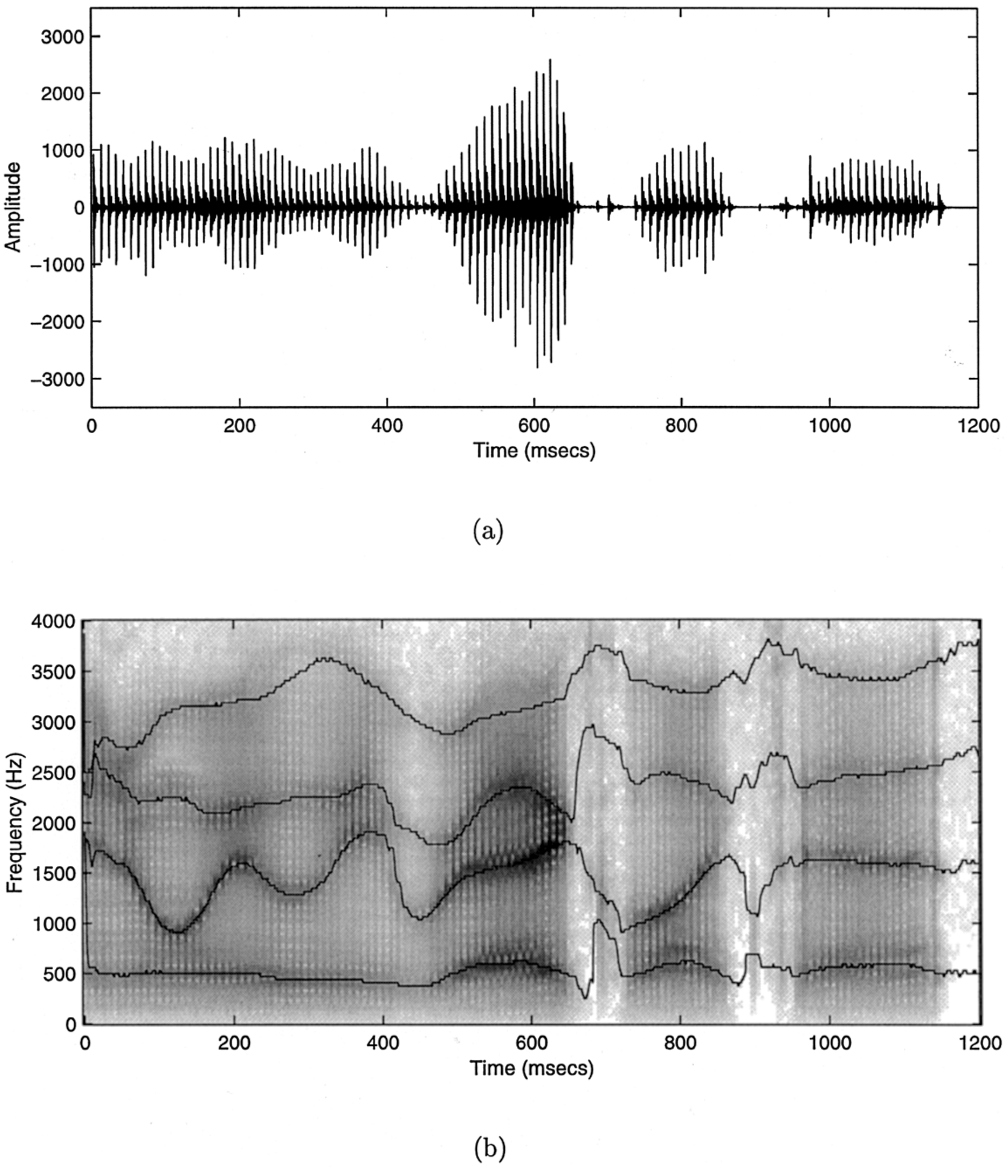
$$s_k(t) = a_0 e^{j\omega_i t} \prod_{i=1}^{P} (1 - p_i e^{j\Omega t}) \prod_{i=1}^{Q} (1 - q_i e^{j\Omega t}) \qquad (3)$$

where $p_1, p_2, \cdots, p_P$ and $q_1, q_2, \cdots, q_Q$ denote the polynomial's roots; $p_i = |p_i| e^{j\theta_i}$ and $q_i = |q_i| e^{j\phi_i}$. p_i denote roots inside the unit circle in the complex plane, and q_i are outside the unit circle. $M = P + Q$. Currently, we assume that there are no roots on the circle, i.e., $|p_i| < 1$ and $|q_i| > 1$. The p_i and q_i are referred to as zeros of the signal $s_k(t)$.

Once $s_k(t)$ is expressed as in (3), the next step is to group the zeros such that the signal may be factored into a MinP part which contains the envelope information and an AllP part which contains nonredundant "phase-only" information. To achieve this grouping, we reflect the zeros that are outside the unit circle (the q_i) to inside the circle (as $1/q_i^*$) and cancel them using poles. Then, we group all the zeros inside the unit circle to form the MinP part of the signal and the zeros outside the circle and the poles that are their reflections inside the unit circle to form the AllP part of the signal. That is

$$s_k(t) = a_0 e^{j\omega_i t} \underbrace{\prod_{i=1}^{P} (1 - p_i e^{j\Omega t}) \prod_{i=1}^{Q} \left(1 - \frac{1}{q_i^*} e^{j\Omega t}\right)}_{\text{MinP}} \cdot \underbrace{\frac{\prod_{i=1}^{Q} (1 - q_i e^{j\Omega t})}{\prod_{i=1}^{Q} \left(1 - \frac{1}{q_i^*} e^{j\Omega t}\right)}}_{\text{AllP}}. \qquad (4)$$

This grouping of signal zeros is analogous to decomposing a linear discrete time system into minimum phase and all-pass systems. Each factor corresponding to a zero or pole in (4) is called an elementary signal [16]. An identity expressing the elementary signals as an infinite series [18], [19] lets us represent



(a)

(b)

Fig. 4. The signal corresponding to the spoken utterance “an oily rag like that” is shown in (a) while (b) displays the estimated DTF’s center frequencies along with the spectrogram.

$s_k(t)$ [given by (4)] compactly as a product of a MinP signal and an AllP signal as follows:

$$s_k(t) = \underbrace{A_c e^{\alpha(t)+\beta(t)+j(\hat{\alpha}(t)+\hat{\beta}(t))}}_{\text{MinP}} \underbrace{e^{j(\omega_c t - 2\hat{\beta}(t))}}_{\text{AllP}} \tag{5}$$

where the “hat” stands for Hilbert transform. Note that for a MinP signal the phase and log-magnitude are related by Hilbert transform. ω_c is $Q\Omega$ plus the arbitrary frequency translation ω_t shown in (3). A_c is $a_0 \prod_{i=1}^{Q}(-q_i)$. The modulation functions associated with $s_k(t)$ are given by [18]–[20]

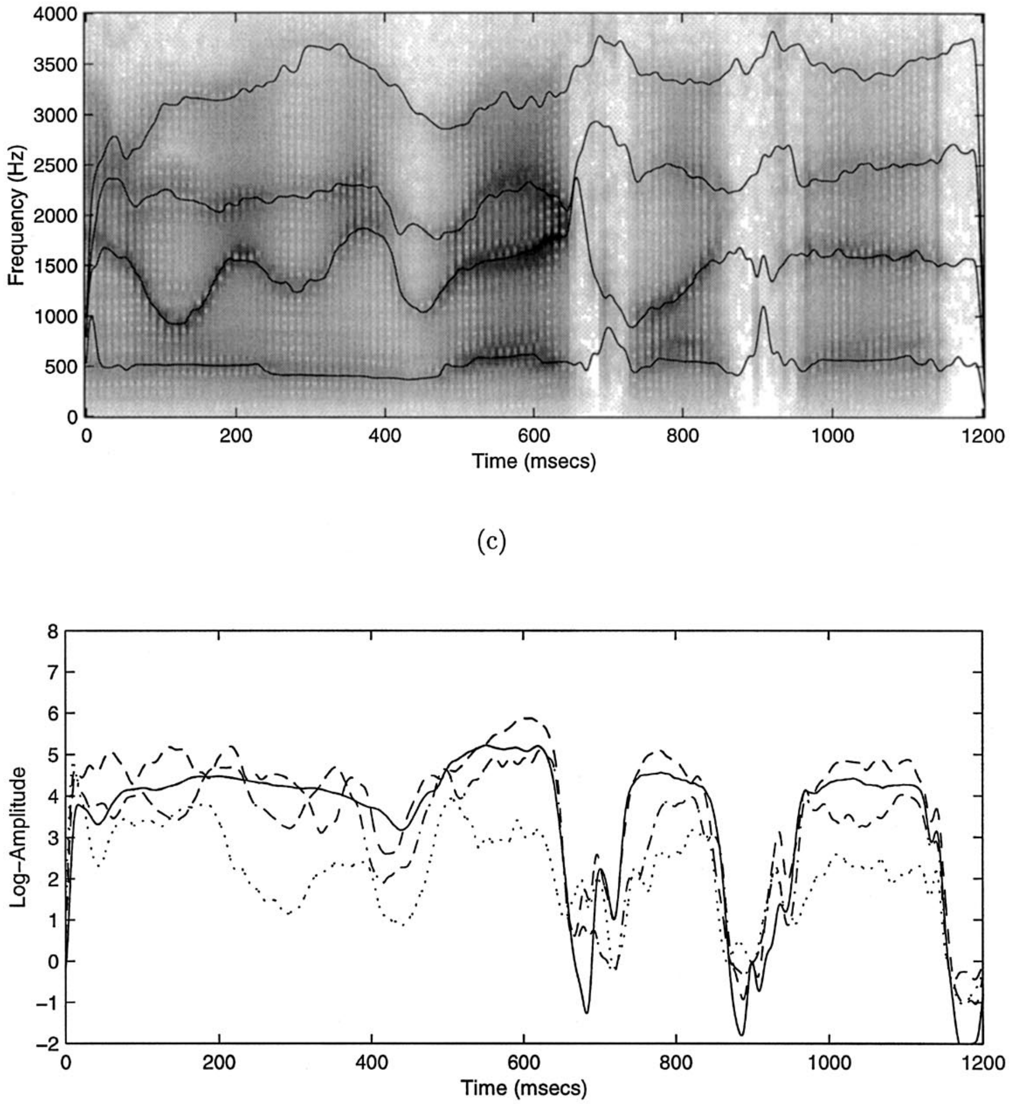
$$\alpha(t) = \sum_{k=1}^{\infty} \sum_{i=1}^{P} -\frac{|p_i|^k}{k} \cos(k\Omega t + k\theta_i) \tag{6}$$

and

$$\beta(t) = \sum_{k=1}^{\infty} \sum_{i=1}^{Q} -\frac{1/|q_i|^k}{k} \cos(k\Omega t + k\phi_i). \tag{7}$$

Closed form expressions can be obtained for $\dot{\hat{\alpha}}(t)$ and $\dot{\hat{\beta}}(t)$ [18], [19]. The “dot” stands for the derivative operation. A detailed description of properties of envelope and IF of signals described by (3) can be found in [19]. We briefly summarize the main points here. The envelope, log-envelope, and phase (or IF) of $s_k(t)$ are not band-limited quantities. It can be shown that $|s_k(t)|^2$ and $(d\angle s_k(t)/dt)|s_k(t)|^2$ (i.e., IF weighted by the square of envelope) are band-limited. It can also be shown that no information is lost by filtering the log-envelope and IF of $s_k(t)$, using a lowpass filter of bandwidth B Hz (recall that we assumed the signal $s_k(t)$ to be the output of a bandpass filter of bandwidth B Hz). That is, in principle, it is possible to essentially reconstruct the signal $s_k(t)$ given ideally filtered versions of the log-envelope and the IF of $s_k(t)$ [19], [20]. Based on this, we propose to represent suitably filtered speech signals using filtered log-envelopes and IF’s.

If $|p_i|$ or $|q_i|$ are close to unity then the modulation functions exhibit wild fluctuations. In traditional AM–FM notation, the



(c)

(d)

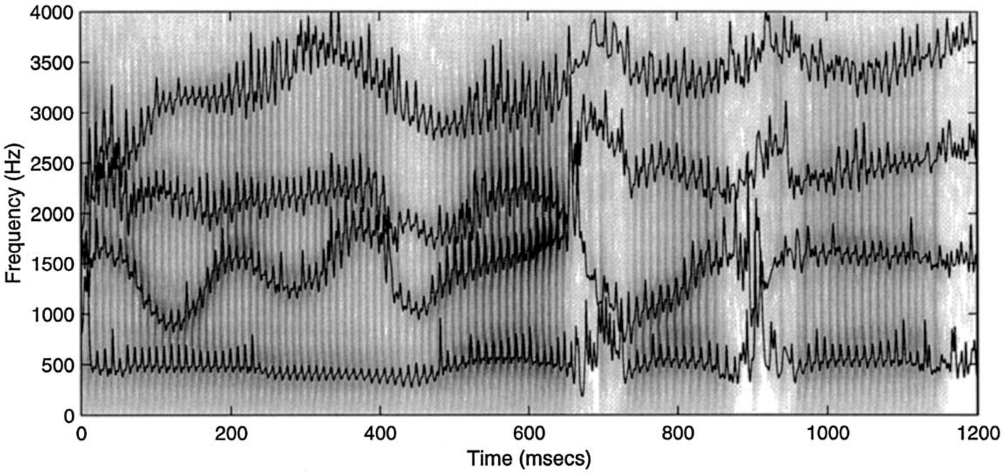
Fig. 4. (*Continued.*) (c) Estimated slowly varying carrier frequencies (overlaid on the spectrogram) and (d) the smooth log-envelopes. As before, solid, dashed, dashed–dotted, and dotted lines correspond to smooth log-envelopes of components at the output of DTF's 1–4, respectively.

AM, which is $|s_k(t)|$, is given by $A_c e^{\alpha(t)+\beta(t)}$ and FM, which is the IF, is $\omega_c + \dot{\alpha}(t) - \dot{\hat{\beta}}(t)$. This IF may take on negative values, which is not physically justifiable. The expression in (5) explicitly shows the Hilbert transform relationship between the log-envelope and phase of the MinP signal. The log-envelope and phase of the MinP signal carry the same information and hence it is sufficient to retain only the envelope. As per the AllP signal component, it can be shown [18] that its IF, i.e., the derivative of its phase function is always positive and always greater than ω_c. This is to be contrasted with traditional IF which can range from positive to negative infinity. If we can separate the MinP and the AllP components of the signal $s_k(t)$, then the envelope and the positive IF (PIF) are uniquely separated and can be used to represent filtered speech. This is the topic of the next section.

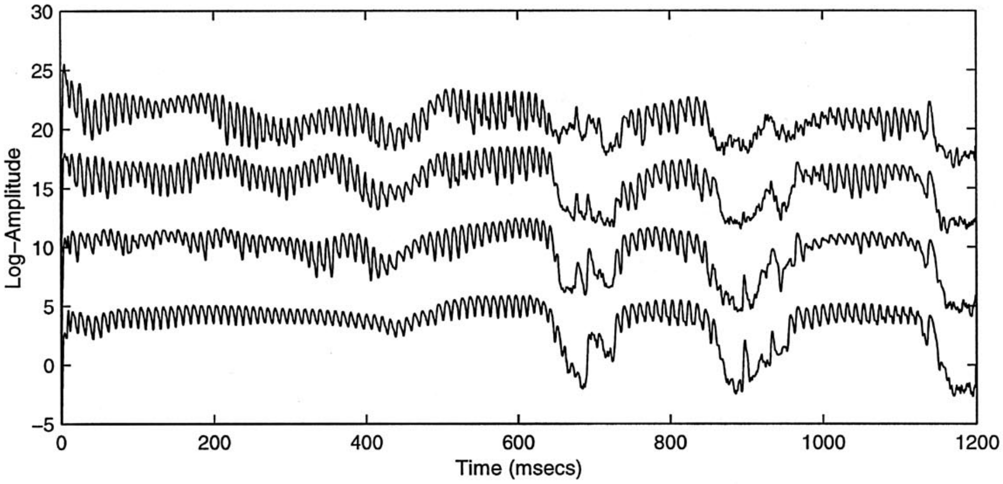
Observe that (5) characterizes $s_k(t)$ only over an assumed processing interval of T s. Since in practice successive overlapping segments of the signal may be processed, the carrier frequency, ω_c, and the carrier amplitude, A_c, are expected to be slowly varying quantities, in comparison to the modulations $\alpha(t)$ and $\beta(t)$. The slowly varying ω_c may be identified with the traditional notion of formant frequency. Given the above signal model, our next step is to decompose the signal $s_k(t)$ into its MinP and AllP component.

III. Separation of MinP-AllP (AM-FM) Components Using Linear Predicton in Spectral Domain

We have indicated in earlier works [18], [19] a number of procedures for decomposing a signal $s_k(t)$ into MinP and AllP components. In this section, we review a simple, elegant algorithm for MinP-AllP decomposition which does not require explicit computation of the logarithm of signal's envelope or rooting of a polynomial. It was called in earlier works [18], [17],



(e)



(f)

Fig. 4. (*Continued*.) In (e) and (f) are shown the PIF's and the log-envelopes, respectively. As before, the PIF's are plotted along with the spectrogram and the log-envelopes are displayed on top of each other, starting with the first DTF's log-envelope at the bottom.

[19] as linear prediction in spectral domain or LPSD. It consists of two parts. In the first part we model the envelope of the signal $s_k(t)$ by minimizing the energy of an error signal $e(t)$, defined as follows:

$$\int_0^T |e(t)|^2\,dt = \int_0^T |s_k(t)h(t)|^2\,dt \qquad (8)$$

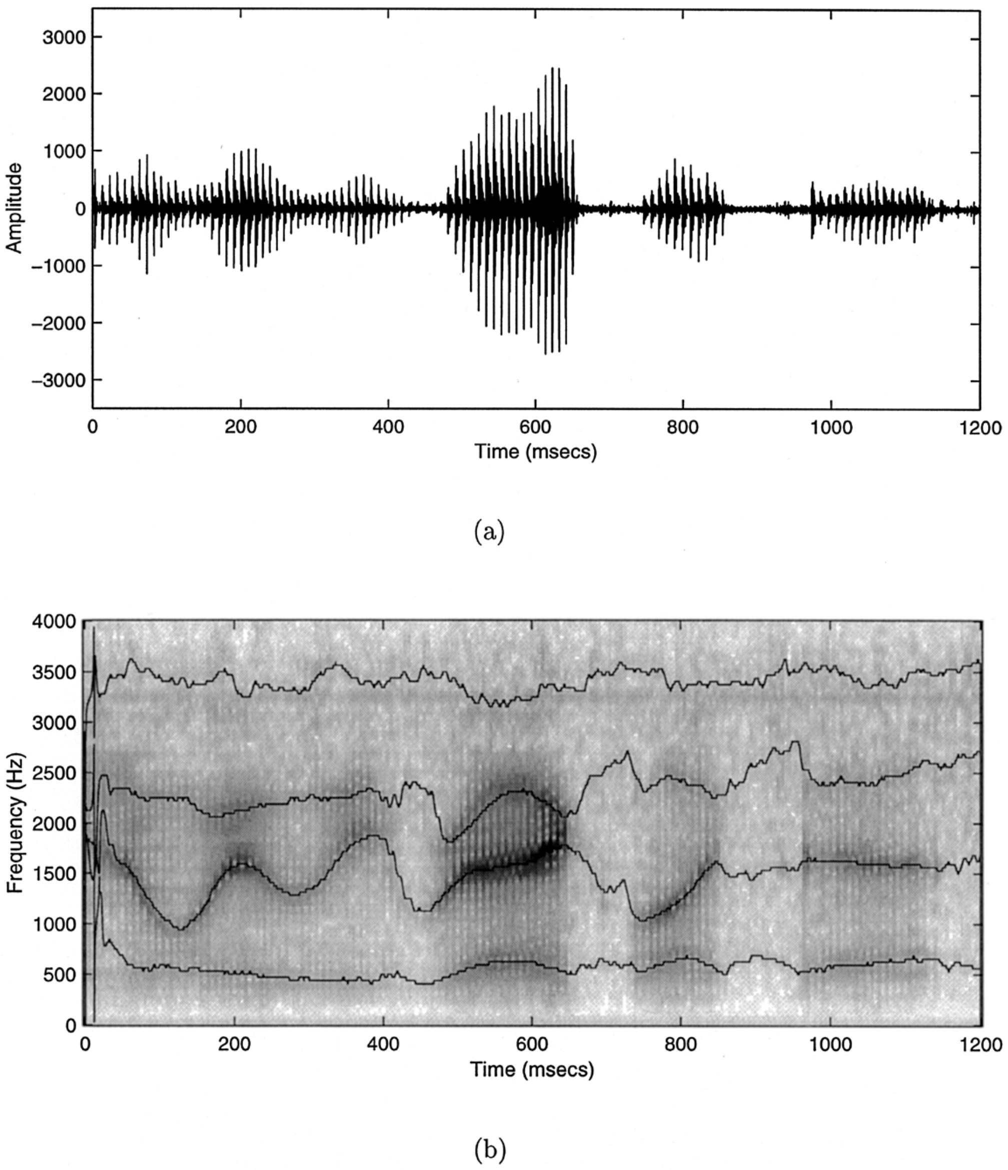
where $h(t) = 1 + \sum_{n=1}^{H} h_n e^{jn\Omega t}$ is an inverse signal. The minimization is achieved by choosing the coefficients h_n. One may recognize this signal envelope modeling method as the analog of the linear prediction (autocorrelation) method well known in spectral analysis [21]. In fact, minimizing the error in (8) amounts to performing linear prediction on the Fourier coefficients of the signal $s_k(t)$ and hence the name LPSD. Similar to the MinP property of the prediction error filter used in linear prediction [21], it can be shown that minimizing $\int_0^T |e(t)|^2\,dt$ will result in an $h(t)$ that is a MinP signal (having all its signal zeros inside the unit-circle). The significance of this property is that $h(t)$'s log-envelope and phase are Hilbert transforms. Because the error minimization is performed to approximate $s_k(t)$'s envelope, if the value of H is chosen sufficiently large, then $h(t)$ will be given by

$$h(t) \approx e^{-(\alpha(t)+\beta(t))} e^{-j(\hat{\alpha}(t)+\hat{\beta}(t))}. \qquad (9)$$

Thus, $1/h(t)$ is the desired approximation to $s_k(t)$'s MinP component. Consequently, the error signal $e(t)$ will be

$$e(t) \approx A_c e^{j(\omega_c t - 2\hat{\beta}(t))} \qquad (10)$$

and, hence, is an approximation to the AllP component of $s_k(t)$. In the second part, shown in Fig. 1 as "measure frequency," the PIF is computed as $\dot{e}(t)/|e(t)|$ or $d\angle e(t)/dt$. Details of the LPSD algorithm using samples of the signal $s_k(t)$ are given in [18] and [19].



(a)

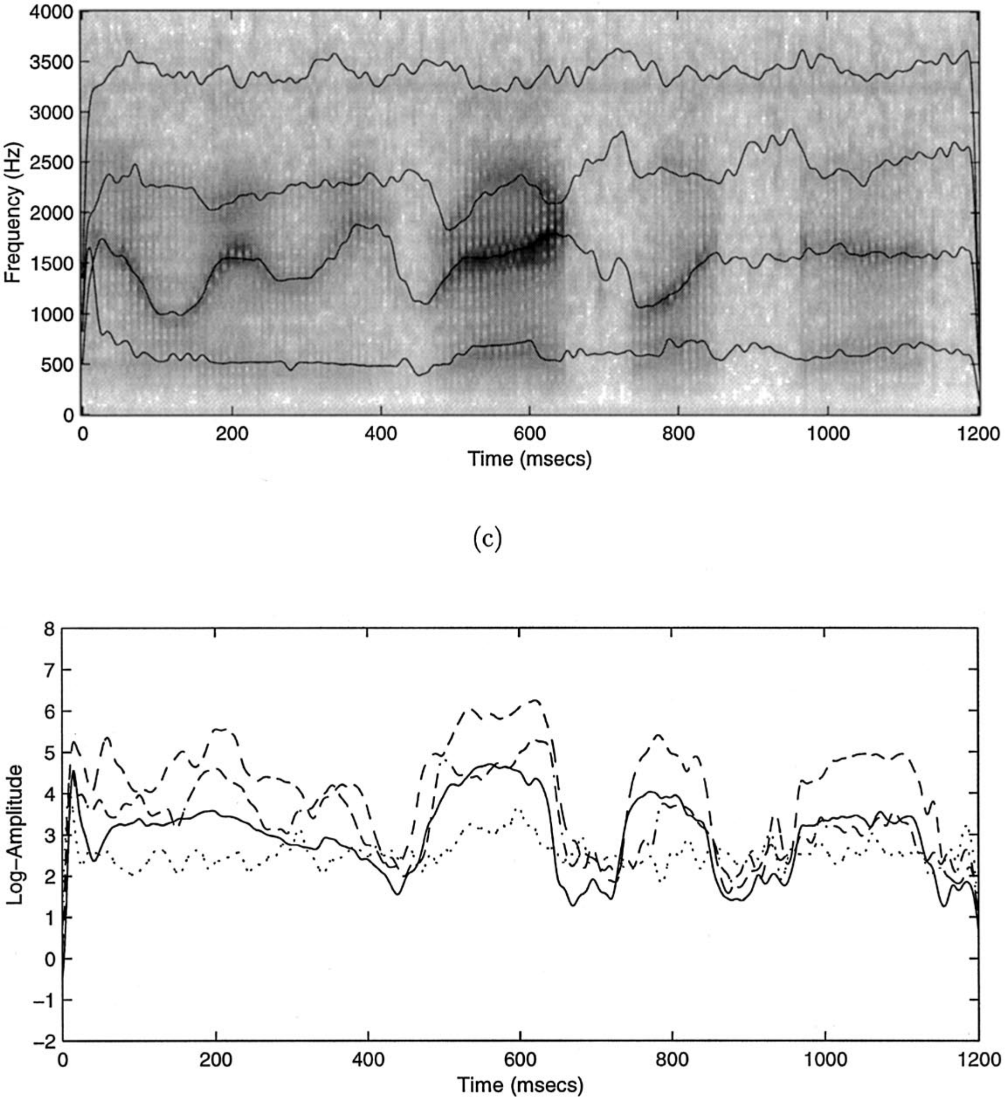
(b)

Fig. 5. The cellular-telephonized version of the utterance "an oily rag like that" considered in Example 2, which is obtained from the CTIMIT database is considered. The signal is plotted in (a). In (b), the four estimated DTFs' center frequencies are displayed.

It may be useful to put into words the purpose of the LPSD algorithm. First, it flattens the envelope of the signal $s_k(t)$ by using an adaptive amplitude demodulator. This is done by multiplying $h(t)$ and $s_k(t)$ and minimizing the resulting error $\int_0^T |e(t)|^2\, dt$. This process not only flattens the envelope of the error $e(t)$, but also removes from the phase of $s_k(t)$ a quantity equal to the Hilbert transform of the log-envelope of $s_k(t)$. This is what causes the IF of $e(t)$ to be positive. Instead if we simply "clip" $s_k(t)$, i.e., obtain $s_k(t)/|s_k(t)|$, then its phase derivative will not have a positive IF. Second, the MinP property of $h(t)$ guarantees that the envelope approximation $1/|h(t)|$ will never equal zero. It is also possible to use the LPSD algorithm to achieve a MinP-MaxP (instead of MinP-AllP) decomposition of $s_k(t)$ [18], [19]. Third, an important advantage of the LPSD algorithm is that it achieves the separation of the MinP-AllP decomposition without explicitly rooting a polynomial or computing the logarithm of the signal $s_k(t)$. For a general time-varying signal like speech, the MinP-AllP decomposition may be achieved by applying the LPSD algorithm over successive windowed portions of the signal and evaluating the resulting modulations at the center of the analysis window.

IV. Time-Varying Front-End Filterbank

Based on discussions in the previous sections, it should be clear that the smoothness of $s_k(t)$'s modulations is inversely proportional to the closeness of its signal-zeros to the unit-circle [20], [19]. The precise locations of the zeros of $s_k(t)$, of course depend on the specific values of the magnitudes and phases of $s_k(t)$'s different spectral components, and hence are functions of the kth filter's parameters and the speech signal $s(t)$. For a signal with a stationary spectrum, a fixed bank of bandpass filters centered around regions with significant spectral content may be employed to decompose the signal into components with smooth modulations. However, for nonstationary signals like speech, such a fixed filter-bank does not seem reasonable.



(c)

(d)

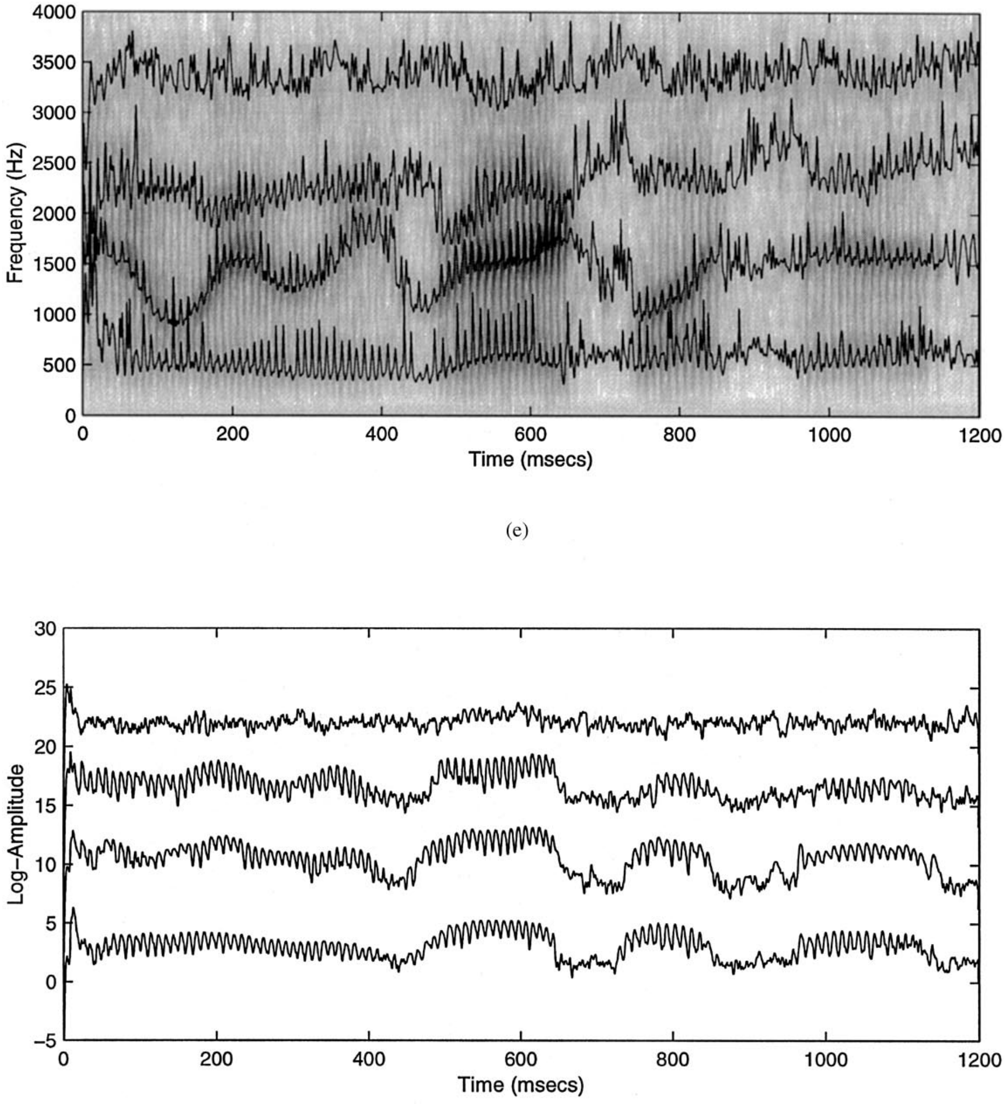
Fig. 5. (*Continued*.) In (c) are shown slowly varying carrier frequencies and (d) log-envelopes. The smooth log-envelopes are plotted using solid, dashed, dashed–dotted, and dotted lines for components at the output of DTF's 1–4, respectively.

Hence, we may require that the front-filters be slowly varying with time such that they are centered roughly around regions with dominant spectral content (or formant frequencies). To do so, we propose an adaptive filter-bank which is an extension of some of our earlier works connected with tracking of narrowband multiple nonstationary sinusoids [28], [31]. The tracking filters that we employ are variations of what are referred to as dynamic tracking filters (DTF's) used for carrier frequency tracking and frequency-feedback demodulation in FM-receivers [32]. In addition, we use all-zero filters (AZF's) in a cross-coupled fashion to suppress interferences from strong neighboring components (formant regions) that otherwise may produce large fluctuations in the modulations of $s_k(t)$ and also significantly alter the estimate of $s_k(t)$'s carrier or center frequency.

Based on the above outlined principle of suppressing neighboring spectral regions with significant energies while tracking (and filtering) a desired spectral region, many variants of our algorithm may be possible. For instance, the DTF may be replaced by any one of the many frequency-locked-loop (or phase-locked-loop) receivers that exist in communication literature [32]. Similarly, the AZF may be replaced by techniques based on residual signal analysis [28], [33] or more generally adaptive Wiener filtering techniques [34]. However, we consider the AZF-DTF combination because it provides flexibility for incorporating (in future works) well-known properties about the auditory periphery like masking, cochlear nonlinearities, and critical bandwidths [29]. The DTF is discussed next.

A. Dynamic Tracking Filter

The DTF's that we employ are single resonance filters which are designed to follow the changing average frequency (or carrier frequency, ω_c) of a signal. Specifically, the complex-valued input signal is filtered using a single pole complex filter, followed by an average frequency or center frequency



(e)

(f)

Fig. 5. (*Continued*.) In (e) and (f), we plot the PIF's and the log-envelopes, respectively. The log-envelopes are stacked on top of each other, just like in previous examples.

estimator. This frequency estimate is then fed back to change the tracking filter's pole-angle. This causes the DTF to follow the slowly varying center frequency of the input signal. Thus, in discrete-time notation, the kth DTF is a time-varying bandpass filter which can be realized by the simple difference equation

$$s_k[n] = r_p \, \exp(j2\pi f_k[n]) s_k[n-1] + (1-r_p)\tilde{s}_k[n] \tag{11}$$

where $\tilde{s}_k[n]$ is the input to the DTF and $s_k[n]$ is its output. r_p is the radius of its single pole [given a bandwidth of B Hz, the pole radius can be computed as $r_p \approx \exp(-B\pi/fs)$, where f_s denotes the sampling frequency], and $f_k[n]$ is the estimated carrier frequency (also referred to as center frequency or average IF) of $s_k[n]$. $1-r_p$ is a constant to ensure unity gain at the DTF's resonant frequency.

B. Algorithm for Filtering Multiple Formants

Now consider the case of multiple formants. Let $f_l[n]$ ($l = 1, 2, \cdots, L$) denote their individual trajectories. Recall that we need to suppress effects due to strong neighboring formants, to represent sinusoidal components contributing to a desired formant. The algorithm we propose uses multiple DTF's, in a cross-coupled configuration that are supplemented by all-zero filters (AZF's) in front of them [31]. The key idea is to adjust zeros of AZF's such that the DTF "sees" essentially components of only one formant. Let the desired formant region be the kth one, with frequency trajectory $f_k[n]$. We want that the kth DTF to track only $f_k[n]$. For this to happen, the kth AZF's output should contain (approximately) components contributing a major portion to the kth formant region. This can be achieved if the kth AZF's zeros are located at $f_l[n]$s ($l = 1, \cdots, L, l \neq k$).

The information about where to place the zeros is derived from all other DTFs' frequency estimates.

Fig. 2 shows the block diagram of one of the channels of the filter bank. The box labeled AZF is the adaptive all zero filter whose $L-1$ zeros track the estimated $f_l[n]$ of all its $L-1$ neighbors. The transfer function of the AZF of the kth tracker at any time index n is

$$H_{Ak}(n,\, z) = K_k[n] \times \prod_{\substack{l=1\\ l\neq k}}^{L} \left(1 - r_z e^{j2\pi f_l[n]} z^{-1}\right) \tag{12}$$

where

$$K_k[n] = \frac{1}{\prod_{\substack{l=1\\ l\neq k}}^{L} \left(1 - r_z e^{j2\pi (f_l[n] - f_k[n])}\right)} \tag{13}$$

ensures unity gain and zero phase lag at frequency $f_k[n]$. Recall that the DTF tracking $f_k[n]$ has a transfer function

$$H_{Dk}(n,\, z) = \frac{1 - r_p}{1 - r_p e^{j2\pi f_u[n]} z^{-1}}. \tag{14}$$

The radius of the AZF's zero, r_z, is set to be slightly less than unity; r_p is the pole radius as before. Ideally, the outputs of individual DTF's consist of components that contribute most to their respective formant regions.

At the output of each DTF, a segment of samples corresponding to an assumed period of T s (typically $1/B$) is considered; appropriate delay is taken into account for causal processing. A single-pole linear prediction (LP) analysis of this segment is then performed. This results in an estimate of the center frequency (or carrier frequency $f_k[n]$) for that filter (i.e., the kth DTF). We emphasize that these $f_k[n]$'s are slowly varying quantities; and hence in practice their estimation and update may be done at a rate much slower than the signals sampling rate. The filtered output is then decomposed according to (5) into MinP and AllP components, using the LPSD algorithm described in the earlier section. The LPSD procedure results in a unique positive AM (of the MinP signal) and an error signal, $e[n]$, with positive IF. In our simulations, we use a two-stage LPSD [18]. In the first-stage, we compute the AllP signal's ($e[n]$'s) IF by phase-angle differencing, i.e., by computing $\angle\{e[n]e^*[n-1]\}$. However to eliminate the possibility of frequency-wrapping [19], [20] we use the absolute mean of this IF's estimate to demodulate $e[n]$, then compute the IF, and add back the constant by which $e[n]$ was demodulated. This IF is then fed once again to the LPSD algorithm (the second stage LPSD). The resulting MinP signal's envelope (from the second-stage LPSD) thus obtained forms the estimate of PIF. Finally, the log-envelopes and PIF's of the components (forming the DTF's outputs) are filtered using a single pole lowpass filter having bandwidth of B Hz consistent with the DTF's bandwidths of B Hz. A Matlab program that implements the above algorithm is provided in [19]. However, the algorithm described in [19] uses a different center-frequency estimator, based on the LPSD algorithm; it performs poorly compared to the center-frequency estimator based on LP analysis mentioned above; further the algorithm as implemented has not been optimized from a computation stand-point.

1) Algorithm's Merits and Demerits: The adaptive filter-bank we have proposed results in a reasonable isolation of formant regions. It is automatic (requires no manual intervention) and is computationally simple, since it incorporates simple filtering techniques. It is stable since the pole-radii of DTF's are fixed. Although this may seem to result in DTF's having fixed bandwidths, the composite responses (considering the cascade of AZF's) result in front-filters with variable bandwidths. The proposed algorithm provides scope for improvements such as, the DTF's bandwidths can be varied based on their center-frequencies (for example to have a constant-Q nature) or based on the log-envelopes' estimates at their outputs. The radii of AZFs' zeros can also be adjusted based on estimates of smooth log-envelopes to incorporate time-varying masking effects. For instance, the radius of a zero at a location $f_l[n]$ may be computed as $((a_l[n]-a_k[n])/50(a_l[n]+a_k[n]))+(49/50)-0.01$; where $a_l[n]$ and $a_k[n]$ are smooth versions of $|s_l[n]|$ and $|s_k[n]|$, respectively. Notice that this estimated radius is such that it varies between 0.95 and 0.99 (corresponding to approximate bandwidth ranges of 130 and 25 Hz, respectively) based on how strong components in the region of $f_l[n]$ interfere with those in $f_k[n]$. Observe that we have not resorted to any decision-based logic as is commonly done in existing formant trackers; such a scheme can be used to improve the algorithm's performance (especially during unvoiced and silence segments). Finally, a real-version of the DTF can be easily implemented that avoids Hilbert transformation [35].

The algorithm's primary drawbacks are as follows. First, the effect of time-varying filtering on the signal is hard to quantify; although the filtered signals when summed up sound just like the original. Second, choice of the algorithm's parameters (which include the number of DTF's and AZF's, their bandwidths, and their processing intervals) introduces variability in the extracted modulations; the parameters can be fixed based on *a priori* knowledge of formant bandwidths or appropriately tuned based on the application. Another problem is initialization of the tracking filters. Although approaches like LPC or fixed filter-bank could address this problem, we performed the following initialization. Initially, we start off with a single DTF (with no AZF) which simply follows the frequency where spectral energy is maximum; the center-frequency estimator is replaced by simply computing the envelope-weighted-IF [19]. A second DTF is then supplemented by only one AZF whose zero follows the first DTF's center frequency. Thus, in general, the kth DTF has zeros corresponding to only the first $k-1$ DTFs' center frequencies. This initialization procedure can also be used to estimate the number of trackers required, or in other words the number of formants, by measuring the signal energy at the output of each stage. Finally, the problem of "switching of tracks" (which generally does not happen) is resolved by sorting the estimated center frequencies of the DTF's.

V. Results of Computer Simulations

We analyze speech signals obtained from the TIMIT and the CTIMIT databases. All the simulations are performed in the

Matlab environment. We consider three examples. The first two examples represent a female and a male speaker from the TIMIT database, respectively. The third example is from the CTIMIT database and is the cellular-telephonized version of the second example.

Example 1: The segment considered is that of a female speaker from the TIMIT database: `timit/train/dr3/fcke0/si1111.wav`. The original waveform is 18 330 samples long and is sampled at 16 kHz. It corresponds to the spoken utterance "How do we define it." The signal is first decimated by a factor of two (thus the sampling frequency is $f_s = 8$ kHz). It is then pre-emphasized using a filter with a transfer function $1 - 0.98z^{-1}$. We then consider samples 1000–8000. This signal is displayed in Fig. 3(a). Its analytic version is then formed using the FFT-based Hilbert transformer. The resulting complex signal is then fed to the proposed algorithm. The algorithm parameters are as follows. The number of DTF's is set to 4. The DTF's pole-radii are chosen to be 0.9 [corresponding to an approximate 3 dB bandwidth of $B = (\log(0.9) f_s/\pi) = 270$ Hz]; note that the single-pole filters that we use do not provide high stop-band attenuations. All AZF's zeros' radii are initially set to 0.98. The processing interval for all DTF's are calculated as $5f_s/B \approx 150$ samples; hence initialization is performed for 150 samples corresponding to 18.75 ms. Recall that the choice of 150 samples for the processing interval implies that at each time-instant the past 150 samples of the DTFs' outputs are used for carrying out the LP and the LPSD analysis. The LPSD parameter, i.e., its order is chosen to be 12 samples for the first stage and four samples for the second stage. In Fig. 3(b) we plot the estimated center frequencies of the DTF's along with the spectrogram; observe that they closely follow the formant frequencies. The estimated average PIF's are plotted in Fig. 3(c) along with the spectrogram. In Fig. 3(d), we show the estimated average log-envelopes as solid, dashed, dashed–dotted and dotted lines of the components forming the outputs of the DTF's. Fig. 3(e) and (f) display the positive IF's and the log-envelopes estimated using the LPSD procedure; in Fig. 3(e) the PIF's are shown along with the spectrogram; observe that they are quasiperiodic in the voiced-speech segment.

Example 2: The next segment we consider is a male speaker from the TIMIT database: waveform `timit/test/dr1/mreb0/sa2.wav`. Specifically, we consider samples 17 086–36 580 which correspond to the spoken utterance "an oily rag like that" sampled at 16 kHz. It is decimated to a sampling frequency of 8 kHz. The signal is then first pre-emphasized as before. The signal is displayed in Fig. 4(a). Its analytic version is formed as in Example 1. The signal then forms the input to the algorithm. All the algorithm's parameters are same as in the first example. The modulations of the decomposed components are plotted in Fig. 4(b)–(d).

Example 3: In this example, we provide results of analyzing the CTIMIT version of Example 2. Specifically, we consider the speech waveform `ctimit/test/dr1/mreb0/sa2.wav`, which is the cellular-telephonized version of the original second example above; the signal is shown in Fig. 5(a). Using exactly the same algorithm parameters, we present the results in Fig. 5(b)–(f).

VI. Discussion and Conclusion

We do not model speech as a sum of AM–FM components, but simply decompose the waveform around regions of dominant spectral energies and extract modulations. The primary difference between our approach and traditional LPC/cepstral analysis [3] is that we explore the logarithm of a signal, as viewed through a suitable filter, in the time domain instead of signal's Fourier transform domain. Specifically, in our speech decomposition scheme, we capture the slowly varying gross details in the signals spectrum (or formants) using a LPC like analysis followed by a LP analysis in the spectral domain that displays finer time-localized details of the signal in the form of positive modulations. Observe that a signal's logarithm (actually log-derivative) yields a physically meaningful quantity, namely the instantaneous frequency. Further, our motivation for time-domain logarithmic processing is derived from a study of the auditory periphery, as opposed to LPC/cepstral analysis which is motivated by the vocal-tract models. In our opinion, resorting to modulations (in the form of log-envelopes and IF's) for characterizing signals consisting of many sinusoidal components, such as a speech formant, may shed new light on the nature of speech signals and on speaker-specific information.

In addition, we speculate that the modulations may themselves be used as features in applications dealing with computer processing of speech. For instance, the modulations may be down-sampled (presumably at the standard 20 ms frame-rate) to construct feature-vectors as follows. The slowly varying carrier frequencies and log-envelopes may be down-sampled in a fairly simple manner. However, the details in the PIF's and log-envelopes may be processed to extract slowly-varying dominant spectral components in them (which may be subsequently downsampled) similar to the modulation-spectrum and RASTA [36] related ideas. Further, only modulations associated with reasonably high energies (or envelopes) may be explicitly used. For example if a particular DTF's envelope, $a_k[n]$, falls below a certain threshold then we may assume that the confidence in that DTF's center frequency's estimate ($f_k[n]$) is low and hence we may replace $f_k[n]$ by an estimate computed by predicting the same based on past estimates (i.e., $f_k[n-1]$, $f_k[n-2]$, $\cdots$, and so on). Speculating further, one may envision averaging groups of these modulations, where the grouping is performed based on some specific criterion (for instance cross-correlation among modulations of different DTF's). Such features may then be used in isolation or in conjunction with standard cepstral (including cepstral-derivatives) templates [37], [38]. In addition, we foresee that the modulations may be potentially useful in speech applications that use prosodic features [39]. We also hasten to point out that extracting modulations from speech will provide more flexibility in addressing problems like channel-equalization, energy normalization, and noise that are typically faced by present day speech/speaker recognizers. For instance, based on certain known statistics about the modulations, certain spectral bands may be separately analyzed (or even discarded) prior to forming features.

Analyzing speech (or signals in general) using AM–FM signal models has a long history. Although it has been revisited many times, moreso recently, many questions still remain. Some frequently asked questions include: What is AM and FM for naturally occurring (not man-made) signals like speech? How are phase or instantaneous frequency (FM) and envelope (AM) related to each other? Is FM more important than AM? Can one guarantee an arbitrary signal's FM to be positive? and so on. Further, one of the factors that has discouraged speech researchers in resorting to AM-FM modeling is: If bandlimited speech (at the output of a bandpass filter) is decomposed into AM and FM, then the resulting modulations are, rather ironically, wild in the sense that their bandwidths are much greater (at times tending to infinity) compared to that of the original bandlimited signal. Finally, a natural question that often arises in speech processing is the need for a front-end filterbank: should one resort to an FFT-based filterbank, or is Mel-filterbank a better choice, or is there an optimal (in some sense) filterbank? In some of our previous works [17]–[20], we have addressed some of these issues. In this paper, we extended these ideas to decompose speech signals. Thus, the difference between previous AM–FM modeling approaches and the work reported here is that we have proposed a methodical way of decomposing speech into modulated components.

The AM and FM contributions of an analytic signal that we consider are such that they convey independent information. If we consider the MinP-AllP representation, the MinP part conveys the AM information, i.e., $e^{\alpha(t)+\beta(t)}$ [or equivalently its logarithm $\alpha(t)+\beta(t)$] around the carrier ω_c; the AllP part conveys the positive FM information in the signal, i.e., $\omega_c - 2\dot{\hat{\beta}}(t)$ which is obtained after removing the redundant phase information due to the MinP part $[(\hat{\alpha}(t) + \hat{\beta}(t)]$ from the signal's phase. Similarly, a MinP/MaxP representation can be used, in which case the signal may be represented by a positive AM [i.e., the envelope of its MinP part, $e^{\alpha(t)}$] and by a positive IF of its MaxP part, i.e., $\omega_c - \dot{\hat{\beta}}(t)$.

There are other interesting aspects in our decomposition. First, MinP signals will have their energy concentrated over a relatively small region in the spectral domain. Second, these are unique signals in that all other signals having identical envelopes are bounded by the phases of MinP (and maximum-phase) signals [16]. Third, since logarithms of MinP signals are themselves analytic they could be further decomposed into their respective slowly varying carrier frequencies and modulated components leading to a tree-like decomposition. These issues are being currently investigated. We believe that the research presented in this paper will stir interest in the direction of taking a new-look at a seemingly old problem of speech analysis. Further, recent results [40] indicate that the envelope and PIF can be represented by certain zero-crossings.

References

[1] 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.

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

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

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-4, pp. 357–366, 1980.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[6] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Processing*, vol. 2, Oct. 1994.

[7] B. S. Atal, "Automatic speech recognition: A communication perspective," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, May 1999, pp. 457–460.

[8] H. Dudley, "The carrier nature of speech," *Bell Syst. Tech. J.*, vol. 19, pp. 495–515, 1940.

[9] E. C. Cherry and V. J. Phillips, "Some possible uses of single sideband signals in formant-tracking systems," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1067–1077, 1961.

[10] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, pp. 412–419, 1980.

[11] B. S. Atal and C. H. Shadle, "Decomposing speech into formants: A new look at an old problem," *J. Acoust. Soc. Amer.*, vol. 1, p. S162, Fall 1978.

[12] J. Flanagan, "Automatic extraction of formant frequencies from continuous speech," *J. Acoust. Soc. Amer.*, vol. 28, pp. 110–118, 1956.

[13] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Energy separation in signal modulations with applications to speech," *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[14] L. B. Jackson and J. Bertrand, "An adaptive inverse digital filter for formant analysis of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 1976, pp. 84–86.

[15] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[16] H. B. Voelcker, "Toward a unified theory of modulation—Part I: Phase-envelope relationships," *Proc. IEEE*, vol. 54, no. 3, pp. 340–354, 1966.

[17] R. Kumaresan and A. Rao, "Unique positive FM–AM decomposition of signals," *Multidimen. Syst. Signal Process.*, vol. 9, pp. 411–418, 1998.

[18] ——, "Model based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1912–1924, March 1999.

[19] A. Rao, "Signal analysis using product expansions inspired by the auditory periphery," Ph.D. dissertation, Univ. Rhode Island, Kingston, June 1997.

[20] A. Rao, "On instantaneous frequency of multicomponent signals," Ph.D. dissertation, Univ. Rhode Island, Kingston, June 1997.

[21] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[22] J. Flanagan and R. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, 1966.

[23] P. Hedelin, "A representation of speech with partials," in *The Representation of Speech in the Peripheral Auditory System*, R. Carlson and B. Granström, Eds, New York: Elsevier, 1982, pp. 247–250.

[24] L. B. Almeida and J. M Tribolet, "Nonstationary spectral modeling of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, pp. 664–678, June 1983.

[25] R. J. McAulay and T. F. Quatieri, "Speech analysis—Synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 744–754, 1986.

[26] R. Kumaresan and C. S. Ramalingam, "On separating voiced-speech into its components," in *Proc. 27th Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Nov. 1993, pp. 1041–1046.

[27] C. S. Ramalingam and R. Kumaresan, "Voiced-speech analysis based on the residual interfering signal canceler RISC algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Adelaide, Australia, Apr. 1994, pp. 473–476.

[28] R. Kumaresan, C. S. Ramalingam, and A. Rao, "RISC: An improved Costas estimator-predictor filter-band for decomposing multi-component signals," in *Proc. 7th Statistical Signal Array Processing Workshop*, Quèbec City, Quèbec, P.Q., Canada, June 1994, pp. 207–210.

[29] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 2nd ed. ed. London, U.K.: Academic, 1988.

[30] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM–FM separation using auditory motivated filters," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 465–480, 1997.

[31] A. Rao and R. Kumaresan, "Dynamic tracking filters for decomposing nonstationary sinusoidal signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Detroit, MI, May 1995.

[32] S. Haykin, *Communication Systems*, 2nd ed, New York: Wiley, 1987.

[33] J. P. Costas, "Residual signal analysis," *Proc. IEEE*, vol. 68, pp. 1351–1352, Oct. 1980.

[34] S. Haykin, *Adaptive Filter Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[35] C. S. Ramalingam, A. Rao, and R. Kumaresan, "Time-frequency analysis using the residual interfering signal canceler filter bank," in *Proc. IEEE-SP Int. Symp. Time-Frequency, Time-Scale Analysis*, Philadelphia, PA, Oct. 1994, pp. 500–503.

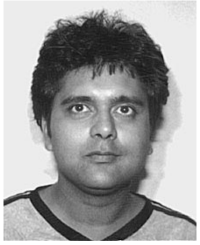
[36] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, San Francisco, CA, Mar. 1992, pp. 121–124.

[37] K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Seattle, WA, 1998, pp. 617–620.

[38] J. Holmes *et al.*, "Using formant frequencies in speech recognition," in *Proc. 5th European Conf. Speech Communication Technology*, vol. 4, 1997, pp. 2083–2086.

[39] E. Shriberg *et al.*, "A prosody-only decision-tree model for disfluency detection," *Proc. Eurospeech*, vol. 5, pp. 2383–2386, 1997.

[40] R. Kumaresan and Y. Wang, "A new real-zero conversion alogrithm," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Istanbul, Turkey, June 2000.



Ashwin Rao (S'94–M'97) was born in Bombay, India. He received the B.S. degree in electrical engineering from the University of Bombay, and the M.S. and Ph.D. degrees from the University of Rhode Island, Kingston.

He spent the Summer of 1995 and the Fall of 1996 as a Graduate Research Assistant with the Department of Neuroscience, Brown University, Providence, RI. In 1996, he was a Summer Intern with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ. In 1997, he served as a Postdoctoral Research Associate with the Department of Physics and Astronomy, Hunter College, City University of New York. Since October 1997, he has been a Research Scientist with Dragon Systems, Inc., Newton, MA, conducting research in the area of large vocabulary continuous speech recognition. His research interests include digital signal processing, speech recognition, cochlear mechanics, and wireless communications.

Dr. Rao is an associate member of the Acoustical Society of America.

Ramdas Kumaresan (S'78–M'79–SM'91–F'93) received the B.S. degree in electronics and communication engineering from the University of Madras, Madras, India, and the M.S. and Ph.D. degrees from the University of Rhode Island, Kingston, in 1979 and 1982, respectively.

He is currently a Professor with the Department of Electrical and Computer Engineering, University of Rhode Island.

Dr. Kumaresan is a Member of the Acoustical Society of America, the Association for Research in Otolaryngology, and the Society for Neuroscience. He held an Alexander von Humboldt Fellowship in 1990–1991 while on sabbatical at the University of Kaiserslautern, Germany.