SIGNAL PROCESSING USING STATISTICAL NONLINEAR SPEECH PRODUCTION MODELS

Wan-Chieh Pai and Peter C. Doerschuk

School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907-1285 USA
{pai, doerschu}@ecn.purdue.edu

ABSTRACT

A nonlinear statistical speech production model based on AM-FM modulation and signal processing methods to extract the component signals are described. Preliminary ideas on using these signals to compute features for a Hidden Markov Model speech recognizer are presented.

1. INTRODUCTION

In this paper we describe a nonlinear speech production model, signal processing tools for the extraction of the information bearing subsignals from the speech signal, and preliminary ideas for the use of these subsignals to compute features for a Hidden Markov Model speech recognizer. The basic idea is that the measured signal is modeled as a superposition of subsignals and each subsignal is a jointly amplitude- and frequency-modulated signal. Therefore, in continuous time, the measured signal $s(t)$ is represented as

$$s(t) = \sum_{i=1}^{I} a_i(t) \cos\left(2\pi \int_{-\infty}^{t} f_i(\tau) \mathrm{d}\tau\right) \qquad (1)$$

where $a_i(t)$ is the ith amplitude message and $f_i(t)$ is the ith frequency message. The goal of the signal processing is to extract $a_i(t)$ and $f_i(t)$ $(i = 1, \ldots, I)$ from the measured signal.

The problem as stated is not well posed because the choice $f_i(t) = 0$ for $i = 1, \ldots, I$, $a_i(t) = 0$ for $i = 2, \ldots, I$, and $a_1(t) = s(t)$ provides an exact representation of $s(t)$. Therefore, we need to introduce additional information about the $a_i(t)$ and $f_i(t)$.

The natural goal of this approach is to represent a broad band signal $s(t)$ in terms of narrow band signals $a_i(t)$ and $f_i(t)$. Even if $s(t)$ does not have a sharply peaked spectrum, this type of representation is promising because the nonlinearity will cause bandwidth expansion exactly as occurs in traditional frequency modulation.

The approach we pursue to extracting $a_i(t)$ and $f_i(t)$ from the data [1] is to introduce statistical models for $a_i(t)$ and $f_i(t)$. In addition, corresponding to the approximation problem, we hypothesize that the measured data is $s(t) + v(t)$ where $v(t)$ is a noise with known statistical properties. Then, under appropriate assumptions, the goal of reconstructing $a_i(t)$ and $f_i(t)$ from $s(t) + v(t)$ can be posed as a Bayesian estimation problem and solved using nonlinear filtering ideas. We call this statistical nonlinear filtering approach the *Model-Based Demodulation Algorithm* (MBDA).

2. NOTATION

Expectation is denoted by "E". If x is a random sequence then $m_x(k) \doteq E[x(k)]$, $R_x(k_1, k_2) \doteq E[x(k_1)x(k_2)]$, and $P_x(k_1, k_2) \doteq E[(x(k_1) - m_x(k_1))(x(k_2) - m_x(k_2))]$. Independent and identically distributed is abbreviated by i.i.d. The Gaussian probability density function (pdf) with mean m and covariance Λ is denoted by $\mathcal{N}(m, \Lambda)$. The notation "$x \sim p$" means that the random variable (RV) x is distributed according to the pdf p. Transpose is denoted ^{T}. The Kronecker delta function is denoted by δ_{k_1,k_2}.

3. THE MODEL AND SPEECH

There has been extensive recent interest in taking a speech signal $s(t)$ and extracting amplitude $a(t)$ and phase $\phi(t)$ modulations, i.e., $y(t) = a(t)\cos(\phi(t))$, using Teager's energy operator [2, 3, 4, 5, 6, 7, 8, 9]. Both the case of a linear superposition of terms [6], i.e., $y(t) = \sum_i a_i(t)\cos(\phi_i(t))$, and a single term observed in the presence of noise [2] have been investigated. In both cases, the signal is first passed through a bank of filters and then the energy operator is applied to the output of each filter. In the case of a superposition of terms, the bandwidth of the ith filter is determined by the bandwidth of the term $a_i(t)\cos(\phi_i(t))$ and the outputs of the ith energy operator are $a_i(t)$ and

$\phi_i(t)$. Therefore, each filter is responsible for a particular term. In the case of a single term in the presence of noise, the bandwidths of the filters are determined by the trade-off between suppressing the noise and passing as much signal energy as possible and the single signal is tracked (by an energy measure) as it moves from filter to filter.

As described in Section 1, in the MBDA approach we simultaneously consider a linear superposition of terms and the presence of noise. In a qualitative sense, the nonlinear filter acts as a bank of bandpass filters where the center frequency of the ith filter tracks the instantaneous frequency of the $a_i(t)\cos(\phi_i(t))$ term and the bandwidth of the ith filter is set to achieve the optimal trade-off between passing signal energy and rejecting noise based on the statistical model. In this point of view, the parameters of the energy operator approach, specifically the bandwidth and center frequencies of the Gabor filters, are seen to qualitatively correspond to the parameters in the statistical model of MBDA.

In speech, a discrete-time version of the model of Section 1 is more natural and the frequency f_i is split into two parts: a slowly-varying center frequency denoted by $f_i(k)$ (the formant frequency with variation on time-scales greater than the pitch period) and a rapidly-varying message frequency denoted by $\nu_i(k)$ (the instantaneous frequency with variation on time-scales shorter than the pitch period). In addition, for each term in the linear superposition of subsignals, there is an instantaneous amplitude signal, denoted by $a_i(k)$, and a total phase signal, denoted by $\phi_i(k)$. If detailed statistical knowledge concerning $a_i(k)$, $f_i(k)$, and $\nu_i(k)$ is available, then it can be incorporated into the mathematical model. However, in the speech application, only rather imprecise information about power and bandwidth is available. Therefore, we have chosen simple dynamics: The instantaneous amplitude and instantaneous frequency signals a_i and ν_i are modeled as first-order autoregressive (AR) processes which allows independent control of the power and the bandwidth. The formant frequency f_i is modeled as a random walk. This choice was made because we expect the formant frequency to remain constant over periods of milliseconds in duration and a random walk is the only Gauss-Markov model in which such behavior has a large probability of occurring. The dynamics of the total phase signal $\phi(k)$ are completely determined by its definition: $\phi_i(k) = \phi_i(0) + 2\pi T\sum_{l=0}^{k-1}(f_i(l) + \nu_i(l))$ where T is the sampling interval. The measured signal, denoted by $y(k)$, is the linear superposition of the contribution from each formant, specifically, $a_i(k)\cos(\phi_i(k))$, plus additive measurement noise. The complete model is therefore

$$
\begin{aligned}
a_i(k+1) &= \alpha_{a_i} a_i(k) + q_{a_i} w_{a_i}(k) && (2)\\
\nu_i(k+1) &= \alpha_{\nu_i} \nu_i(k) + q_{\nu_i} w_{\nu_i}(k) && (3)\\
f_i(k+1) &= f_i(k) + q_{f_i} w_{f_i}(k) && (4)\\
\phi_i(k+1) &= \phi_i(k) + 2\pi T f_i(k) + 2\pi T \nu_i(k) && (5)\\
y(k) &= \sum_{i=1}^{I} a_i(k)\cos(\phi_i(k)) + rv(k) && (6)
\end{aligned}
$$

where the process noises w_{a_i}, w_{ν_i}, and w_{f_i} and the observation noise v are all iid $N(0,1)$ sequences; the initial conditions are $a_i(0) \sim N(0, q_{a_i}^2/(1-\alpha_{a_i}^2))$, $\nu_i(0) \sim N(0, q_{\nu_i}^2/(1-\alpha_{\nu_i}^2))$, $f_i(0) \sim N(m_{f_i,0}, p_{f_i,0}^2)$, and $\phi_i(0) \sim N(0, p_{\phi_i,0}^2)$; and the process noises, observation noise, and initial conditions are all independent. Notice that the initial conditions require that $|\alpha_{a_i}| < 1$ and $|\alpha_{\nu_i}| < 1$ (since otherwise the stated variances are negative) in which case a_i and ν_i are wide sense stationary random sequences. For later convenience, define $\theta = (\alpha_{a_i},\ q_{a_i},\ \alpha_{\nu_i},\ q_{\nu_i},\ q_{f_i},\ r,\ m_{f_i,0},\ p_{f_i,0},\ p_{\phi_i,0})$. We estimate the parameter vector θ by matching the second order statistics of the model to training data.

There are several generalizations of the model specified by Eqs. 2–6 that are of interest. The models for $a_i(k)$ and $\nu_i(k)$ in Eqs. 2–6 are first order models and therefore have broad spectra that only rolloff gradually. However, one objective of AM-FM models is to model a relatively broad-band signal by nonlinearly combining quite narrow-band signals. Therefore, narrow-band models, specifically, models with more narrow-band behavior than first order models, are of interest. One choice that includes such narrow-band models is ARMA models, e.g., $\sum_{j=0}^{p}\alpha_{a_i}(j)a_i(k-j) = \sum_{j=0}^{q}\beta_{a_i}(j)w_{a_i}(k-j)$ ($\alpha_{a_i}(0) = 1$). These models can easily be fit into the state-space framework of Eqs. 2–6: Assuming $q < p$, define $\mathbf{b}_i(k) \in \mathcal{R}^p$,

$$
\begin{aligned}
\mathbf{g} &= [1, 0, \ldots, 0]^T \in \mathcal{R}^p,\\
\boldsymbol{\beta}_{a_i} &= [\beta_{a_i}(0), \ldots, \beta_{a_i}(q), 0, \ldots, 0]^T \in \mathcal{R}^p,\\
\mathbf{A}_{a_i} &= \begin{bmatrix} -\alpha_{a_i}(1) & -\alpha_{a_i}(2) & \ldots & -\alpha_{a_i}(p) \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix};
\end{aligned}
$$

replace Eq. 2 by $\mathbf{b}_i(k+1) = \mathbf{A}_{a_i}\mathbf{b}_i(k) + \mathbf{g}w_{a_i}(k)$; and replace Eq. 6 by $y(k) = \sum_{i=1}^{I}\boldsymbol{\beta}_{a_i}^T\mathbf{b}_i(k)\cos(\phi_i(k)) + rv(k)$.

A second generalization to model slow variation of a signal is to use piecewise constant or piecewise linear models. This can also be incorporated into the state-space framework of Eqs. 2–6. A piecewise constant model would replace Eq. 4 by $f_i(k+1) = f_i(k) +$

$q_{f_i} w_{f_i}(k)\delta_{k \bmod N,0}$ while a piecewise linear model would replace Eq. 4 by the two equations

$$\begin{aligned} f_i^+(k+1) &= f_i^+(k) + q_{f_i} w_{f_i}(k)\delta_{k \bmod N,0} \\ f_i^-(k+1) &= f_i^-(k) + [f_i^+(k) - f_i^-(k)]\delta_{k \bmod N,0} \end{aligned}$$

and replace Eq. 5 by $\phi_i(k+1) = \phi_i(k) + 2\pi T\{[f_i^+(k) - f_i^-(k)](k \bmod N)/N + f_i^-(k)\} + 2\pi T\nu_i(k)$. In both cases N is the distance over which f_i is constant or linear and might, for instance, equal the frame duration of the recognizer. These models are time varying but that does not introduce substantial additional computation in the nonlinear filters of Section 4.

A third modification, motivated by sinusoidal speech models [10], is to replace the time-varying formant frequencies by a larger number of fixed frequencies. This can also be incorporated into the state-space framework of Eqs. 2–6: delete Eq. 4, replace Eq. 5 by $\phi_i(k+1) = \phi_i(k) + 2\pi T\nu_i(k)$, and replace Eq. 6 by $y(k) = \sum_{i=1}^{I} a_i(k)\cos(2\pi T f_i k + \phi_i(k)) + rv(k)$ where f_i are the fixed frequencies which might be chosen according to $f_i = f_1 + i\Delta$ for constants f_1 and Δ.

4. NONLINEAR FILTERS

If $a_i(k)$ was constant then Eqs. 2–6 describe a frequency modulated communication system, the Extended Kalman Filter (EKF) [11, Section 8.2] is essentially a phase-locked loop (PLL), and the PLL is an excellent estimator. Therefore, we compute the estimates $\hat{a}_i(k|k)$, $\hat{\nu}_i(k|k)$, $\hat{f}_i(k|k)$, and $\hat{\phi}_i(k|k)$ (hereafter, we will not indicate the conditioning which is always $k|k$) by using the EKF for this more complicated model. The computational requirements are minimal: the state equation is already linear, the one-step state transition matrix (denoted by F) is block diagonal (1 block per formant) and each block is sparse so multiplication by F is inexpensive, and the observation is a scalar so the one matrix inversion is actually division by a scalar. The result of the EKF are the estimates $\hat{a}_i(k)$, $\hat{\nu}_i(k)$, $\hat{f}_i(k)$, and $\hat{\phi}_i(k)$. From these estimates we can compute a reconstructed speech signal, denoted by $\hat{y}(k)$, by $\hat{y}(k) = \sum_i \hat{a}_i(k)\cos(\hat{\phi}_i(k))$.

More sophisticated nonlinear filters than the EKF could also be used, e.g., Refs. [12, 11, 13]. However, we have achieved interesting results using the simple EKF and, when using more sophisticated filters, problems of robustness and failure of the model statistics to match the data statistics become more severe.

5. SPEECH EXAMPLE

In this section we describe the application of these ideas to the sentence "Even then, if she took one step forward he could catch her." from the TIMIT database [14, dr1/fcjc0/si1207]. The model has 4 formants with initial conditions $m_{f_i,0}$ of 500, 2000, 2900, 4100 Hz for $i = 1$, 2, 3, and 4 respectively. For all 4 formants, $\alpha_{a_i} = \alpha_{\nu_i} = .99$, $p_{f_i,0} = 0$, and $p_{\phi_i,0} = 0$. The values of q_{a_i}, q_{ν_i}, and q_{f_i} vary from formant to formant: $q_{a_i} = \sqrt{1700}, \sqrt{120}, \sqrt{220}, \sqrt{3000}$; $q_{\nu_i} = \sqrt{61}, 7, 10, 10$; and $q_{f_i} = \sqrt{1.5}, \sqrt{1.5}, \sqrt{1.5}, 1$ for $i = 1, 2, 3, 4$ respectively. Finally, $r = \sqrt{1/12}$. The spectrogram of the original speech with superimposed plots of the estimates $\hat{f}_i(k)$ is shown in Figure 1(a). [The spectrogram is computed by dividing the signal into 8 ms frames (each contains 128 samples) with 4 ms (64 sample) overlap between adjacent frames and then computing the magnitude (in dB) of the 128 point FFT of each frame]. In Figures 1(a,b), the formant tracks extend through regions of the spectrogram where there is little energy because at sample k we plot the ith formant track $\hat{f}_i(k)$ even when the energy in the ith formant (essentially the energy in $a_i(k)$) is small. The spectrogram of $\hat{y}(k)$ (i.e., the speech reconstructed from the EKF outputs) is shown in Figure 1(b) and closely matches the speech spectrogram shown in Figure 1(a). Figures 1(a,b) demonstrate the smooth behavior of the the EKF estimates and the accurate reconstruction of the speech in unvoiced regions even though the model used by the EKF is really a model for voiced speech.

6. SPEECH RECOGNITION

In the Hidden Markov Model (HMM) approach to speech recognition, the initial step is to transform the speech signal into a sequence of feature vectors. By removing aspects of the speech signal that are irrelevant for recognition, this step achieves data compression and simplifies the estimation of conditional measurement probability densities in the HMM. If the AM-FM modulation model for speech production describes physical behavior that is missing in linear speech production models, then features based on the AM-FM modulation model may improve the performance of HMM recognizers.

A natural first step toward using AM-FM features is to determine how the AM-FM model can be used to generate features analogous to standard features. Two important classes of standard features are spectral features derived from filter banks and from linear predictive coding (LPC) [15] and here we focus on filter banks. Let $s(k)$ be the input speech signal and $H_j(z)$ be the filters in the bank which are FIR linear phase filters with center frequencies f_j and bandwidths W_j and which approximately satisfy $\sum_{i=j}^{J} H_j(\exp(j\Omega)) = 1$. We pass $s(k)$ through H_j to generate $y_j(k)$. We trans-

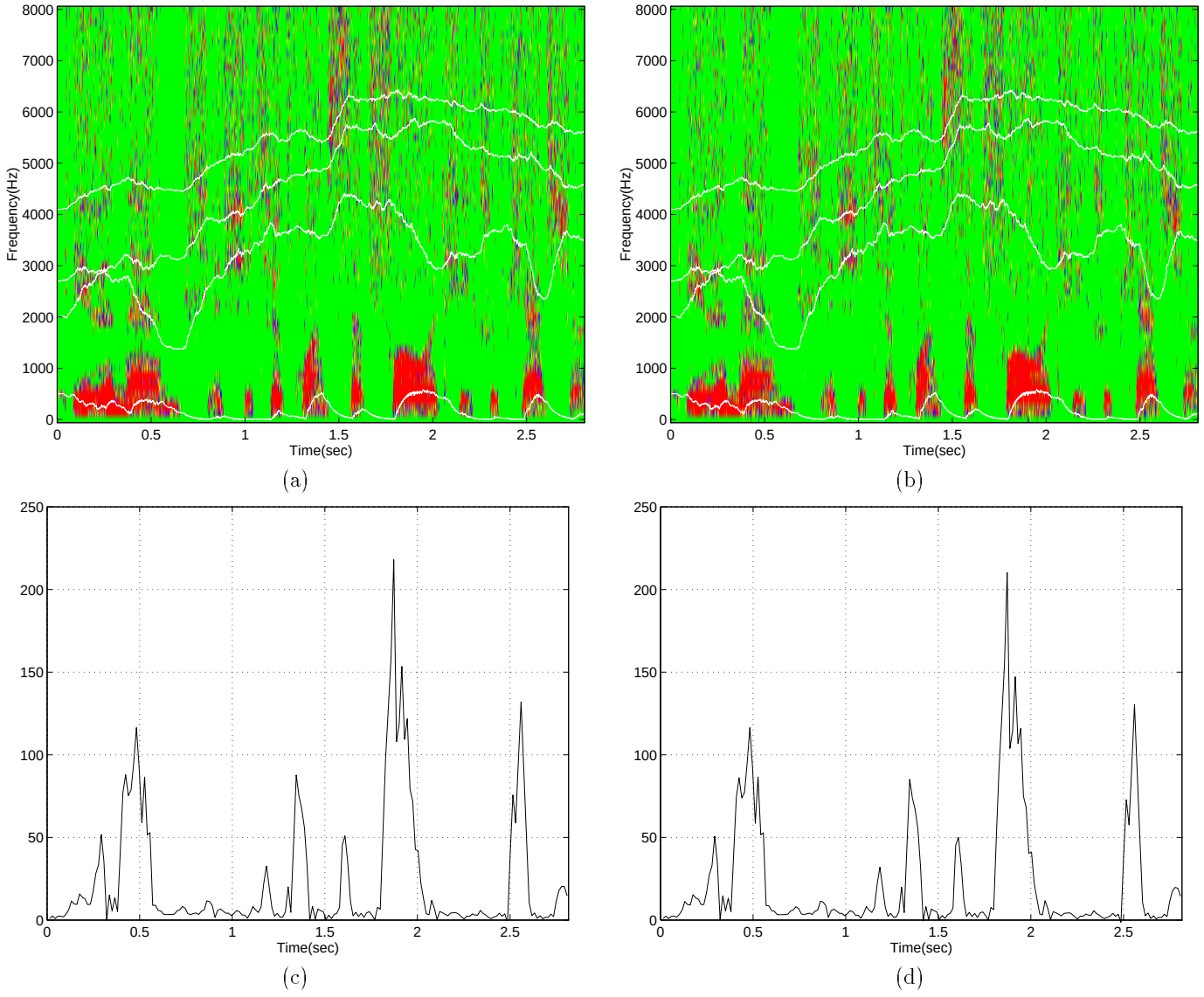

Figure 1: The sentence "Even then, if she took one step forward he could catch her." (a) Original spectrogram and estimated formant tracks. (b) Reconstructed spectrogram and estimated formant tracks. (c) Standard filter bank features. (d) EKF-based features.

form from passband to baseband by low pass filtering $\sqrt{2}y_j(k)\cos(2\pi f_j Tk)$. We subsample the result based on the bandwidth W_j to give a set of signals $u_j(k)$. At least approximately it is possible to recover the speech signal $s(k)$ from the signals $u_j(k)$. Traditional filter bank features, denoted by $\zeta_j(k)$, are computed by taking the absolute value of $u_j(k)$, low pass filtering, and subsampling to the frame rate where the bandwidth of the low pass filter is chosen so that the frame rate is the Nyquist rate. New features analogous to filter bank features and denoted by $\eta_j(k)$ are computed in two steps. First, apply an EKF (based on a 1 formant model with f_1 fixed at 0) to the signals $u_j(k)$, to estimate $\hat{a}_j(k)$, $\hat{\phi}_j(k)$, and $\hat{\nu}_j(k)$. Then apply exactly the same processing used to transform $u_j(k)$ to $\zeta_j(k)$ for traditional features to transform $\hat{a}_j(k)$ to $\eta_j(k)$ for new features. The reason we compute $u_j(k)$ signals in a way that makes it possible to at least approximately reconstruct $s(k)$ is that the AM-FM model is supposed to be a speech production model so it is most natural to use it on signals from which speech can be reconstructed. The reason for applying the EKF to $u_j(k)$ rather than the speech itself is that it requires less computation: in order to get narrow band estimates we would have to generalize the model of Eqs. 2–6 with some of the ideas in the final paragraphs of Section 3 and a J-formant model of that type requires significant computation.

In Figures 1(c,d) we show the $j = 3$ components of traditional and new features for a system with $J = 9$; 80th order linear phase FIR bandpass filters with passbands of [100, 400], [400, 700], [700, 1000], [1000, 1350], [1350, 1800], [1800, 2400], [2400, 3280], [3280, 4690], [4690, 7150]; a 300 Hz bandwidth for the first low pass filter; a subsampling by 1/26 to compute $u_j(k)$; a 33.3 Hz bandwidth for the second low pass filter; a subsampling by 1/9 to compute $\zeta_j(k)$; and an EKF with parameters $q_a = 1$, $q_\nu = 9$, $\alpha_a = \alpha_\nu = 0.99$, and $r = 0.2$. The plots are time-shifted to remove the delay introduced by the linear phase FIR filters. With this set of EKF parameters, the two sets of features (Figures 1(c,d)) are very similar because the EKF is constructing the signal primarily by varying the amplitude $a(k)$ with only small variation of the phase $\phi(k)$. If the bandwidth of $a(k)$ in the model used by the EKF is decreased, then the EKF constructs the signal with more equal variation in the amplitude $a(k)$ and the phase $\phi(k)$.

Similar processing applied to $\hat{\phi}_j(k) - 2\pi f_j Tk$ will yield new features, which are phase-sensitive features, based on the AM-FM model.

7. REFERENCES

[1] Shan Lu and Peter C. Doerschuk. Nonlinear modeling and processing of speech based on sums of AM-FM formant models. *IEEE Trans. Sig. Proc.*, 44(4):773–782, April 1996.

[2] Alan C. Bovik, Petros Maragos, and Thomas F. Quatieri. AM-FM energy detection and separation in noise using multiband energy operators. *IEEE Trans. Sig. Proc.*, 41(12):3245–3265, December 1993.

[3] Robert B. Dunn, Thomas F. Quatieri, and James F. Kaiser. Detection of transient signals using the energy operator. In *Proc. IEEE ICASSP-93*, volume III, pages 145–148, 1993.

[4] Helen M. Hanson, Petros Maragos, and Alexandros Potamianos. Finding speech formants and modulations via energy separation: With application to a vocoder. In *Proc. IEEE ICASSP-93*, volume II, pages 716–719, 1993.

[5] James F. Kaiser. Some useful properties of Teager's energy operators. In *Proc. IEEE ICASSP-93*, volume III, pages 149–152, 1993.

[6] Petros Maragos, James F. Kaiser, and Thomas F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. Sig. Proc.*, 41(10):3024–3051, October 1993.

[7] Petros Maragos, James F. Kaiser, and Thomas F. Quatieri. On amplitude and frequency demodulation using energy operators. *IEEE Trans. Sig. Proc.*, 41(4):1532–1550, April 1993.

[8] Herbert M. Teager. Some observations on oral air flow during phonation. *IEEE Trans. ASSP*, 28(5):599–601, October 1980.

[9] Alexandros Potamianos and Petros Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. In *Proc. IEEE ICASSP-95*, pages 784–787, 1995.

[10] Robert J. McAulay and Thomas F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. ASSP*, 34(4):744–754, August 1986.

[11] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1979.

[12] D. R. Polk and S. C. Gupta. Quasi-optimum digital phase-locked loops. *IEEE Trans. Commun.*, 21(1):75–82, January 1973.

[13] Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

[14] W.M. Fisher, V. Zue, J. Bernstein, and D. Pallett. An accoustic-phonetic database. In *113th Meeting of the Acoustical Society of America*, 1987.

[15] Lawrence Rabiner and Juang Biing-Hwang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.