MODELING AND PROCESSING SPEECH WITH SUMS OF AM-FM FORMANT MODELS

Shan Lu and Peter C. Doerschuk

School of Electrical Engineering, Purdue University
West Lafayette, IN 47907–1285 USA
lu@ecn.purdue.edu doerschu@ecn.purdue.edu

ABSTRACT

We describe a new approach to decomposing signals that are modeled as a sum of jointly amplitude and frequency modulated cosines with slowing-varying center frequencies observed in noise based on statistical nonlinear filtering ideas. We demonstrate the ideas on a formant tracking problem for the sentence "Where were you while we were away."

1. INTRODUCTION

There has been extensive recent interest in taking a signal $y(t)$ 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, 5, 6, 7, 8, 9]. Both the case of a linear superposition of terms [7], 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 i th filter is determined by the bandwidth of the term $a_i(t)\cos(\phi_i(t))$ and the outputs of the i th 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.

In our approach, which we call the *Model-Based Demodulation Algorithm* (MBDA), we simultaneously consider a linear superposition of terms and the presence of noise. We describe the signal in terms of statistical models for a_i, ϕ_i, and the noise and apply nonlinear filtering techniques to estimate a_i and ϕ_i from the noisy signal. In a qualitative sense, the nonlinear filter acts as a bank of bandpass filters where the center frequency of the i th filter tracks the instantaneous frequency of the $a_i(t)\cos(\phi_i(t))$ term and the bandwidth of the i th 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.

This work was supported by U. S. National Science Foundation grant MIP-9110919 and a Whirlpool Faculty Fellowship.

2. MODEL

For each formant (i labels the formant), there is a dynamical system which describes the time evolution of 4 signals: the Kaiser-Teager amplitude signal ($a_i(k)$), the Kaiser-Teager frequency signal ($\nu_i(k)$), the formant frequency ($f_i(k)$), and the total phase signal ($\phi_i(k)$). We have chosen simple dynamics: The Kaiser-Teager amplitude and 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. Mathematically, this is reflected in the fact that if $x(k)$ is a random walk then $E[x(k)]$ is constant and $x(k) = \arg\max_{x(k+1)} p(x(k+1)|x(k))$. 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

$$\begin{aligned}
a_i(k+1) &= \alpha_{a_i} a_i(k) + q_{a_i} w_{a_i}(k) && (1)\\
\nu_i(k+1) &= \alpha_{\nu_i} \nu_i(k) + q_{\nu_i} w_{\nu_i}(k) && (2)\\
f_i(k+1) &= f_i(k) + q_{f_i} w_{f_i}(k) && (3)\\
\phi_i(k+1) &= \phi_i(k) + 2\pi T f_i(k) + 2\pi T \nu_i(k) && (4)\\
y(k) &= \sum_i a_i(k)\cos(\phi_i(k)) + r v(k) && (5)
\end{aligned}$$

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.

3. NONLINEAR FILTERS

The goal of the nonlinear filtering problem is to estimate $a_i(\cdot)$, $\nu_i(\cdot)$, $f_i(\cdot)$, and $\phi_i(\cdot)$ from the measurements $y(\cdot)$ and knowledge of the statistical model described above. If $a_i(k)$ was constant then Eqs. 1–5 describe a frequency modulated communication system, the Extended Kalman Filter (EKF) [1, Section 8.2] is essentially a phase-locked loop (PLL), and the PLL is an excellent estimator. Therefore, we compute estimates of $a_i(k)$, $\nu_i(k)$, $f_i(k)$, and $\phi_i(k)$ based on the measurements $y(0), \ldots, y(k)$ by using the EKF for this more complicated model. The EKF operates by linearizing the nonlinear model around the current best estimate and then applying the Kalman Filter (KF) to the resulting linear time-varying 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. It is necessary, however, to compute trigonometric functions at each time step in order to linearize the observation equation.

The result of the EKF are estimates of $a_i(k)$, $\nu_i(k)$, $f_i(k)$, and $\phi_i(k)$, which are denoted $\hat{a}_i(k)$, $\hat{\nu}_i(k)$, $\hat{f}_i(k)$, and $\hat{\phi}_i(k)$ respectively. From these estimates we can compute a reconstructed speech signal, denoted by $\hat{y}(k)$, by $\hat{y}(k) = \sum_i \hat{a}_i(k)\cos(\hat{\phi}_i(k))$

4. SYNTHETIC EXAMPLE

In this example, patterned after [8, Fig. 3], we consider a chirp signal:

$$y(k) = \cos(2\pi f_a kT)\cos(2\pi(f_m + f_c kT)kT)$$

where $T = 1/16000$ s, $f_a = 30$ Hz, $f_m = 500$ Hz $f_c = 2000$ Hz/s, and k is in the range from 1 to 1600 (i.e., 100 ms). Therefore, the instantaneous frequency in the signal is $f_m + 2f_c kT$ while the instantaneous frequency in our model Eqs. 1–5 is $f(k) + \nu(k)$. The results, shown in Figures 1 and 2, are excellent: after an initial transient, the filter accurately tracks the increasing formant frequency $f(k)$, the zero Kaiser-Teager frequency $\nu(k)$, and the oscillating amplitude $a(k)$. (The EKF used one formant with $\alpha_a = .99$, $q_a = .1$, $\alpha_\nu = .99$, $q_\nu = .1$, $q_f = 3$, $r = \sqrt{1/12}$, $m_{f,0} = f_m$, $p_{f,0} = 0$, and $p_{\phi,0} = 0$).

5. FORMANT TRACKING EXAMPLE

The model (Eqs. 1–5) and EKF can be applied to many different problems in speech processing. In this section we emphasize the formant signal $\hat{f}$ rather than the Kaiser-Teager frequency signal ν or amplitude signal a and apply the model and EKF to the problem of tracking the formants through an entire sentence. The sentence is "Where were you while we were away." from the TIMIT database [4, dr1/msjs1/sx9]. The model has 4 formants with initial conditions $m_{f_i,0}$ of 450, 1300, 2000, and 3100 Hz for $i = 1, 2, 3$, and 4 respectively. For all 4 formants, $\alpha_{a_i} = \alpha_{\nu_i} = .99$, $q_{\nu_i} = 12$, $p_{f_i,0} = 0$, and $p_{\phi_i,0} = 0$. The values of q_{a_i} and q_{f_i} vary from formant to formant: $q_{a_i} = 50, 30, 10, 1$; and $q_{\nu_i} = \sqrt{.5}, \sqrt{22}, 2, 2$ 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 3(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 Figure 3(a), the formant tracks extend through regions of the spectrogram where there is little energy because at sample k we plot the i^{th} formant track $\hat{f}_i(k)$ even when the energy in the i^{th} formant (essentially the energy in $a_i(k)$) is small. Figure 3(a) demonstrates excellent tracking of the formants in this sentence in spite of large and rapid variation in the formant frequencies.

From the EKF outputs we compute the reconstructed speech signal $\hat{y}(k)$. In Figure 3(b) we show the spectrogram of $\hat{y}(k)$ which is very similar to the spectrogram of $y(k)$ shown in Figures 3(a).

6. DISCUSSION

In this paper we describe the MBDA algorithm. We are currently investigating the performance of MBDA for other speech processing problems, such as speech coding, and comparing the performance and computational complexity of MBDA with alternative algorithms, such as DESA-1 [7, Section 5.2].

7. REFERENCES

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

[2] A. C. Bovik, P. Maragos, and T. 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] R. B. Dunn, T. F. Quatieri, and J. F. Kaiser. Detection of transient signals using the energy operator. In *Proc. IEEE ICASSP-93*, pages III–145–III–148, 1993.

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

[5] H. M. Hanson, P. Maragos, and A. Potamianos. Finding speech formants and modulations via energy separation: With application to a vocoder. In *Proc. IEEE ICASSP-93*, pages II–716–II–719, 1993.

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

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

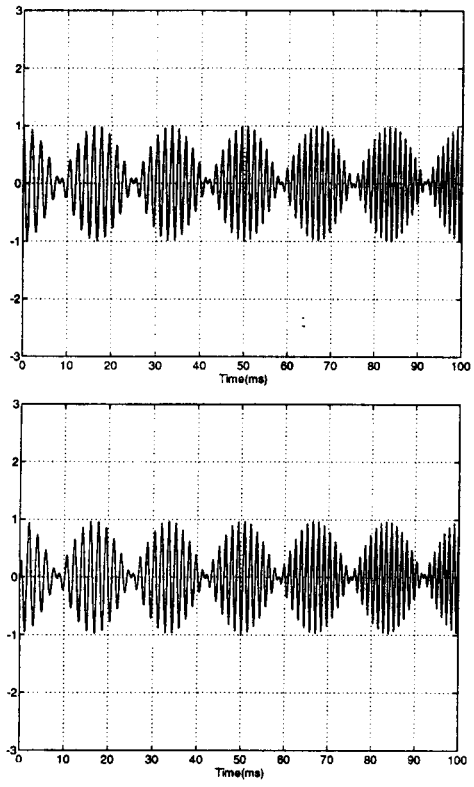


Figure 1: Chirp signal. (a) Original chirp signal. (b) Reconstructed chirp signal.

[8] P. Maragos, J. F. Kaiser, and T. F. Quatieri. On amplitude and frequency demodulation using energy operators. *IEEE Trans. Sig. Proc.*, 41(4):1532–1550, April 1993.

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

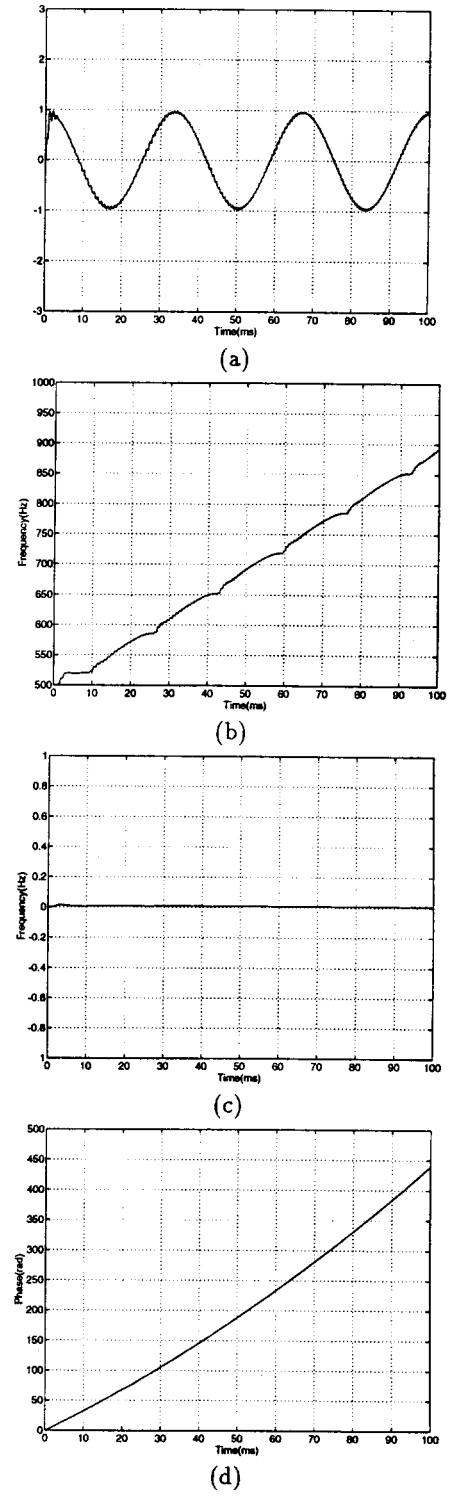
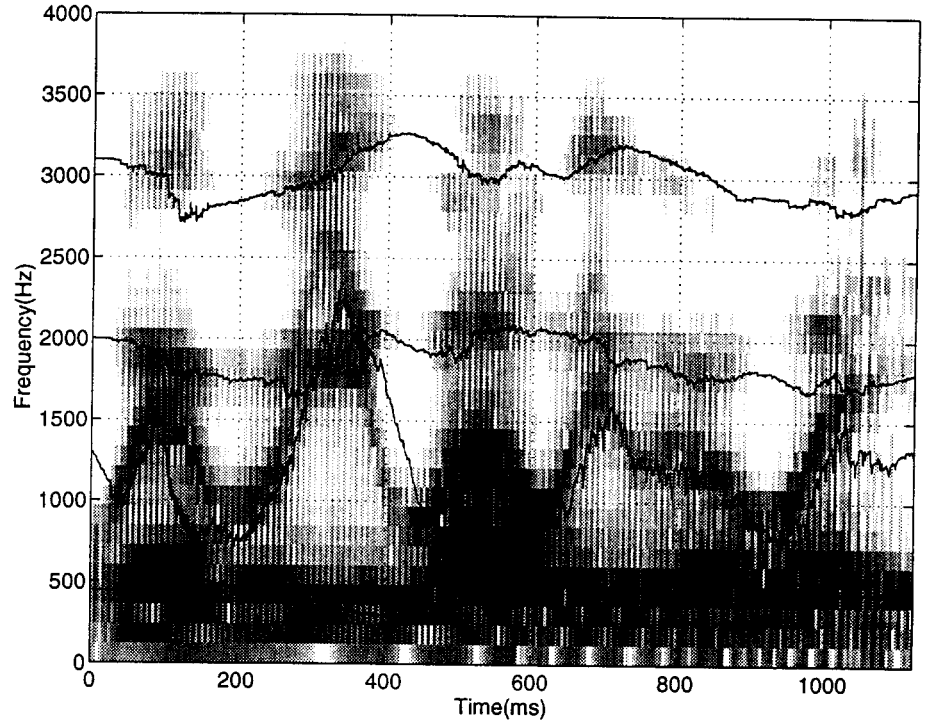
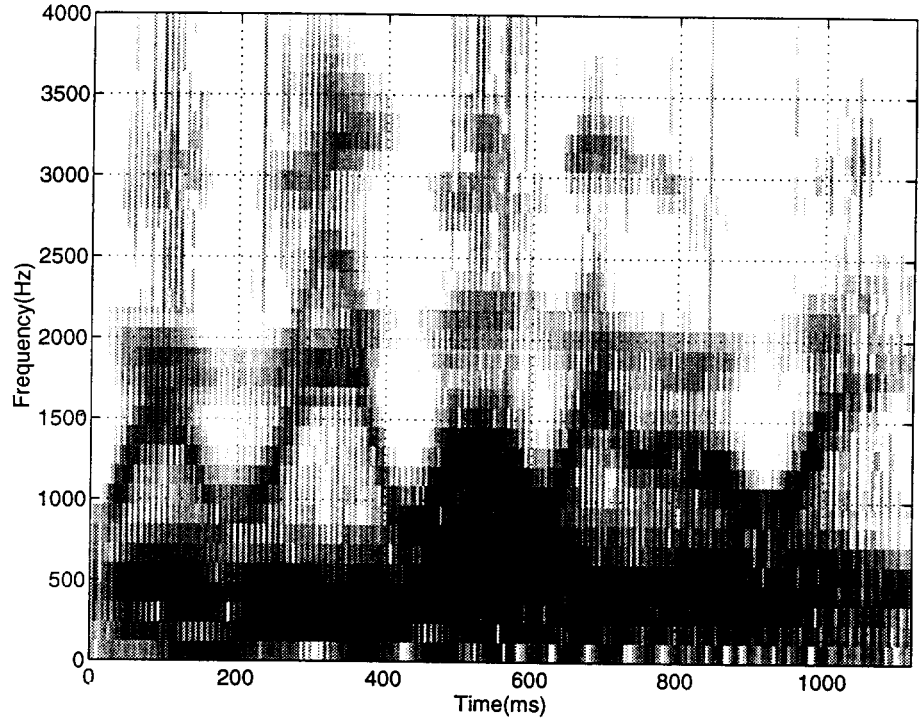


Figure 2: Chirp signal. (a) Estimated amplitude signal $\hat{a}(k)$. (b) Estimated formant frequency signal $\hat{f}(k)$. (c) Estimated Kaiser-Teager frequency signal $\hat{\nu}(k)$. (d) Estimated total phase signal $\hat{\phi}(k)$.



(a)



(b)

Figure 3: Spectrogram and formant tracks for the sentence "Where were you while we were away." (a) Spectrogram of original signal and formant tracks. (b) Spectrogram of reconstructed signal.