10-1-1995

NONLINEAR MODELING AND PROCESSING OF SPEECH WITH APPLICATIONS TO SPEECH CODING

Shan Lu
Purdue University School of Electrical and Computer Engineering

Peter C. Doerschuk
Purdue University School of Electrical and Computer Engineering

Nonlinear Modeling and Processing of Speech With Applications to Speech Coding

Shan Lu
Peter C. Doerschuk

TR-ECE 95-23
October 1995

School of Electrical
and Computer Engineering
Purdue University
West Lafayette, Indiana 47907-1285

NONLINEAR MODELING AND PROCESSING OF SPEECH WITH APPLICATIONS TO SPEECH CODING

Shan Lu and Peter C. Doerschuk[1]

School of Electrical and Computer Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN 47907–1285

[1]This work was supported by a Whirlpool Faculty Fellowship. I!.S. National Science Foundation grant IMIP-9110919, and the School of Electrical and Computer Engineering, Purdue University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In recent years there has been increasing interest in nonlinear speech modeling. In our approach, a speech signal is modeled as a sum of jointly amplitude (AM) and frequency (FM) modulated cosines with slowly-varying center frequencies. The key problem is to extract the center frequency and the amplitude and frequency modulations for each formant in the model from the measured speech signals.

In this study, we describe the speech signal in terms of statistical models and apply statistical nonlinear filtering techniques (Extended Kalman Filter) to estimate the amplitude and frequency. The AM and FM signals are estimated for all the formants simultaneously in an efficient and computationally tractable manner. Using Cramer-Rao bound techniques, we can compare the performance of our computationally feasible estimators relative to the performance of the computationally intractable optimal estimator. Recombination of the amplitude and frequency signals generated by our approach results in faithful reconstruction of speech in both the time and frequency domains.

We consider two applications. The first application, which is formant tracking, is a direct application of our nonlinear filters since the formant frequencies are a part of our nonlinear model. The application of our entire framework to speech coding is also discussed.

1. INTRODUCTION

There has been extensive recent interest in modeling a speech resonance using a signal $y(t)$ with time-varying amplitude $a(t)$ and phase $\phi(t)$, i.e., $y(t) = a(t)\cos(\phi(t))$, where $a(t)$ is an amplitude-modulation (AM) and $\phi(t)$ is a phase-modulation (PM). If $\phi(t)$ is the integral of a more fundamental signal, then PM is really frequency-modulation (FM). The initial motivation for modeling a speech resonance using an AM-PM or AM-FM structure is Teager's work on nonlinear modeling of time-varying speech resonances [1, 2].

In this chapter, we first provide a brief description of speech production mechanism and linear speech modeling ideas. Then we present evidence of nonlinear effects in speech. A nonlinear speech model and an existing demodulation method are introduced next. The potential applications of the nonlinear model is also discussed. Finally we provide an overview of this study and an outline of the technical reporl,.

1.1 Speech Production

Speech is produced by vocal organs which consist of lungs and trachea., larynx, and vocal tract. Lungs supply compressed air to the system which is delivered by way of the trachea. The larynx is a complicated system of cartilages and muscles containing and controlling the vocal cords whose opening and closing can form a. quasi-periodic pulse train. The glottal pulse train, which is the principle excitation source for speech, is then modulated or filtered by the vocal tract.

Acoustically, the vocal tract is a tube of nonuniform cross section, approximately

17 cm long in adult males, which is usually ope11 at one end and nearly closed at the other. Such a tube is a. distributed-parameter structure and thus has many natural frequencies. The term "speech resonances" refers to the oscillator systems formed by local cavities of the vocal tract which emphasize certain frequencies and de-emphasize other frequencies during speech production. These resonances, also known as formants, are the most important acoustical characteristics of the vocal tract. The glottal pulse train is rich in harmonics and these harmonics interact strongly with the vocal tract resonances to affect the tone quality of the voice. Formants thus provide the listener's primary source of information about the position of the speaker's vocal organs [3].

1.2 Linear Speech Models

In linear speech modeling, speech is clescribed by a. linear prediction (LP) model

$$y(k) = a_1 y(k-1) + \cdots + a_p y(k-p) + e(k),$$

where $y(\cdot)$ is the discrete-time speech signal, p is the model order: $a_1, \ldots, a,$ are the prediction coefficients and $e(\cdot)$ is the precliction error.

$y(\cdot)$ can also be viewed as the output of an all-pole linear filter with $a_1, \cdot \ldots a_p$ as the filter coefficients and $e(\cdot)$ as the input. When the order p is properly chosen, the all-pole filter, sometimes referred to as a. vocal tract filter, is a plausible model of the vocal tract. The poles of the linear filter transfer function characterize speech formants. In the linear moclel, the moclel coefficients, and hence the formants. are assumed constant over each short-time analysis frame (about 10-30 ms). Thus this classic approach assumes some local stationarity of the speech signal.

1.3 Nonlinear Effects In Speech

Experimental evidence in Teager's work [1, 2] has motivated researchers [4, 5] to investigate the possibility of relaxing this local stationarity assumption and using a

more refined model where variations of the phase and amplitude of speech resonances can be modeled and detected on an instantaneous-sample time scale.

Teager found evidence that speech resonances exhibit more complicated modulation structure than a linear model could possibly describe. Consider the all-pole linear filter model introduced above. Each pair of complex conjugate poles corresponds to a second-order resonator with an exponentially-damped cosine as its impulse response:

$$h_{lin}(t) = Ae^{-\sigma t}\cos(\omega_c t + \theta), \tag{1.1}$$

where ω_c is the center (formant) frequency and $\sigma > 0$ controls the formant bandwidth. If a signal representing a speech resonance were produced by a second-order linear resonator, which is inferred in linear speech modeling, then the signal would have a exponentially decaying envelope. In contrast, Teager found that bandpass filtering speech vowel signals around formants resulted in signals with several envelope "bumps" per pitch period ([4] Figures 5–7, [1] Figure 5). These bumps indicate some kind of modulation in each formant.

Teager's work has also provided indications and plausible explanations of how the speech resonances can change rapidly both in frequency and amplitude even within a single pitch period, based on rapidly-varying and separated airflows in the vocal tract. It is known that slow time variations of the elements of a simple second-order oscillator can result in amplitude or frequency modulation of the simple oscillator's cosine response. To see this, consider an undriven, undamped oscillator consisting of a mass m and a spring with stiffness coefficient k. The equation of motion is

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \omega_i^2 x = 0, \qquad \omega_i^2 = k/m, \tag{1.2}$$

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

$$\omega_i^2(t) = \omega_c^2[1 + \frac{2\omega_m}{\omega_c}\cos(\omega_f t)] \tag{1.3}$$

If $\omega_m \ll \omega_c$ and $\omega_f \ll \omega_c$, it has been shown [6] that the approximate solution of Eq. (1.2) is

$$x(t) = A\cos[\omega_c t + \frac{\omega_m}{\omega_f}\sin(\omega_f t)], \tag{1.4}$$

which is an FM signal. Similarly se ond-order oscillators with time-varying damping generate responses that contain amplitude modulation [7]. Thus, during speech production, the rapid variation due to separated airflow of air masses and effective cross-.sectional areas of vocal tract cavities can cause modulations of the pressure and velocity fields.

1.4 Nonlinear Speech Model

All these considerations lead to the modeling of a single speech resonance by an AM-FM model [4]

$$y(k) = a(k)\cos(\phi(k)), \tag{1.5}$$

where $\phi(k) \doteq \Omega_c k + \Omega_m \sum_{n=0}^{k} q(n) + \theta$ for some function $q(\cdot)$ and constants Ω_c, Ω_m, and θ and Ω_c is the formant frequency of the resonance. The instantaneous frequency is defined as $\Omega^{\text{inst}}(k) \doteq \Omega_c + \Omega_m q(k)$. The total speech signal y(k) is then modeled as a linear superposition of such AM-FM terms

$$y(k) = \sum_{i=1}^{I} a_i(k)\cos(\phi_i(k)), \tag{1.6}$$

where I is the number of speech formants which are indexed by i.

Obviously, in order to apply the nonlinear speech model to any speech processing problems, it is necessary to estimate the amplitude $a_i(k)$ and phase $\phi_i(k)$ modulations from the measured speech signal $y(k)$. One such estimation algorithm is the energy separation algorithm based on Teager's energy operator [5, 4].

The discrete-time Teager energy operator Ψ, applied to a signal $z(k)$, is defined [4, Eq. (5)] to be $\Psi[z(k)] \doteq z^2(k) - z(k-1)z(k+1)$. (A corresponding continuous-time operator exists but the current study is restricted to discrete-time

problems). Let $y(k)$, representing a single speech resonance, he modeled as in Eq. (1.5). Then the DESA-1 algorithm [4, Eqs. (107)–(108)] for computing estimates of $\Omega^{\text{inst}}(k)$ and $|a(k)|$ from the signal $y(k)$ is defined by the following three equations:

$$z(k) \doteq y(k) - y(k-1) \tag{1.7}$$

$$\widehat{\Omega^{\text{inst}}(k)} \doteq \arccos\left(1 - \frac{\Psi[z(k)] + \Psi[z(k+1)]}{4\Psi[y(k)]}\right) \tag{1.8}$$

$$\widehat{|a(k)|} \doteq \sqrt{\frac{\Psi[y(k)]}{1 - \left(1 - \frac{\Psi[z(k)]+\Psi[z(k+1)]}{4\Psi[y(k)]}\right)^2}} \tag{1.9}$$

where "hat" (i.e., $\hat{}$) indicates an estimate.

Single resonances are extracted from measured speech by bandpass filtering the speech signal with a bank of bandpass filters, such as Gabor filters, with center frequencies at the formant frequencies selected from the short-time speech spectrum [4]. Each filter is responsible for one particular term and the bandwidth of the i th bandpass filter is determined by the bandwidth of the term $a_i(k)\cos(\phi_i(k))$. Thus resonances are assumed to be relatively independent of each other. The energy operator is then applied to the output of each bandpass filter to extract the envelope and instantaneous frequency signals.

In [4], an important issue is the bandwidth of the bandpass filter for extracting speech resonances. It should not be too wide because then significant contributions from neighboring formants will be included. On the other hand, the bandpass filter should not have a very narrow passband because some information in the resonance can be either missed or deemphasized. Methods which optimize the trade-off between these two considerations in choosing the filter bandwidth require additional study.

The case of one resonance observed in noise is considered in [8]. The observed signal is first passed through a bank of bandpass filters. At each instant, the energy operator is applied to the channel response that has the largest energy. The bandwidth 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.

1.5 Application Of Nonlinear Speech Model

The nonlinear speech model can he applied to many speech processing tasks, including speech recognition, speech restoration, and speech coding.

The AM and FM signals extracted from the nonlinear inodel can be incorporated as new features into current speech recognition frameworks, e.g., hidden Markov models (HMM). Since the nonlinear speech model is capable of characterizing rapid variations in speech, incorporating such ideas should generate interesting results. The :nonlinear model can be especially useful in phoneme transitions where the vocal tract changes its shape rapidly and conventional analysis methods based on linear models are insufficient [9, 10, 11, 12].

Speech restoration is another area where the nonlinear model can be useful. The idea. is to estimate AM and FM signals in the presence of a. detailed noise inodel that realistically describes the degraded speech signal, e.g., a model for cockpit noise sources. Then these estimated signals can be combined to yield the restored speech. This approach is particularly promising when the modulations are extracted using statistical estimation methods, since then the design of the algorithms for rejecting noise can be simplified.

If the nonlinear model more accurately reflects physical reality than a. linear model, then coding based on the nonlinear inodel will provide better performance for a given hit rate than coding based on a linear model. One possibility [12] is to adopt techniques similar to the formant [13] and the phase [14] vocoders and combine their itdvantages. Another possibility, which is loosely based upon sinusoidal coding ideas [15, 16, 17], is to incorporate some linear speech coding methods, such a.s LPC and CELP, in the nonlinear speech coder. Of these three broad application areas, we have focused on the speech coding area in this study and our result's are reported in Chapter 5.

1.6 Overview

In this study, we present a novel demodulation algorithm for the AM-FM nonlinear speech model. We describe the signal in terms of statistical models for a_i, ϕ_i, and the noise and apply nonlinear filtering techniques (Extended Kalman Filter) to estimate a_i and ϕ_i from the noisy signal. A linear superposition of terms (i.e., Eq. (1.6)) and the presence of noise are considered simultaneously.

The statistical point of view of our approach simplifies the design of algorithms for the rejection of noise in the speech signal and allows the application of the estensive theory of statistical estimation. The AM and FM signals are estimated for all the formants simultaneously in an efficient, and computationally tractable manner. Using Cramer-Rao bound techniques, we can compare the performance of our computationally feasible estimators relative to the performance of the optimal estimator. Recombination of the amplitude and frequency signals generated by our approach results in faithful reconstruction of speech in both the time and frequency domains.

We consider two applications. The first application, which is formant tracking: is a direct application of our nonlinear filters since the formant frequencies are a part of our nonlinear model. The second application is speech coding. The idea is to use our nonlinear filtering methods to estimate $a_i(k)$ and $\phi_i(k)$ for each formant in the speech signal. These estimates are then coded, transmitted and decoded. Finally, the speech is reconstructed from the decoded estimates. We have experimented with a variety of techniques to code the estimated signals.

The remainder of the technical report is organized as follows: In Chapter 2 we describe the statistical model and estimation problem and the Cramer-Rao bound for the estimation problem. We also discuss parameter identification for the model and a particular suboptimal nonlinear estimator, specifically, the Extended Kalman Filter.. In Chapter **3** we describe the applications of our approach to some synthetic examples and formant tracking problems. In Chapter 4 we compare our approach

with the energy separation algorithm. The application of the entire framework to speech coding is described in Chapter 5. Finally, we summarize our results in Chapter 6.

2. MODEL-BASED DEMODULATION ALGORITHM

Recently there has been substantial interest in taking a signal $y(k)$ and extracting amplitude $a(k)$ and phase $\phi(k)$ modulations using Teager's energy operator [8, 18, 12, 19, 4, 5, 2]. More precisely, the signal is mocleled as $y(k) = a(k)\cos(\phi(k))$ and the goal is to estimate $a(k)$ and $\phi(k)$ (or the first difference of $\phi(k)$) from the measured signal $y(k)$. The purposes of this chapter and next chapter are to propose a corresponding statistical problem formulation, analyze the best-achievable performance for this formulation by computing the Cramer-Rao bound, propose a. practical suboptimal estimator for this formulation, and demonstrate the estimator on several speech analysis applications.

In our approach, which we call the *Model-Based Demodulation Algorithm.* (MBDA), we simultaneously consider a. linear superposition of terms, i.e., $y = \sum_i \mathrm{a}, \cos(\phi_i)$, and the presence of noise. We describe the signal in terms of statistical models for a;, ϕ_i, and the noise and apply nonlinear filtering techniques to estimate a_i and ϕ_i from the noisy signal.

This chapter is organized as follows: In Sections 2.2 and 2.3 we describe the statistical model and estimation problem and the Cramer-Rao bound for the estimation problem. In Sections 2.4 and 2.5 we describe parameter identification for the model and a. particular suboptimal nonlinear estimator, specifically, the Extended Kalman Filter. In Chapter 3 we discuss the application of our approach to some synthetic and real speech problems.

2.1 Notation

Expectation is denoted by "$\mathbf{E}$". If x is a random sequence then $m_x(k) \doteq \mathbf{E}[x(k)]$, $R_x(k_1, k_2) = \mathbf{E}[x(k_1)x(k_2)]$, and $P_x(k_1, k_2) \doteq \mathbf{E}[(x(k_1) - m_x(k_1))(x(k_2) - m_x(k_2))]$. In the case where $R_x(k_1, k_2)$ is a function of only $k_1 - k_2$, the discrete-time Fourier transform of $R_x(0, k)$ is denoted by $S_x(\Omega)$, specifically, $S_x(\Omega) = \sum_{k=-\infty}^{+\infty} R_x(0, k)e^{-j\Omega k}$. Independent and identically distributed is abbreviated by i.i.d. The Gaussian probability density function (pdf) with mean m and covariance A is denoted by $\mathcal{N}(m, 11)$. The notation "$x \sim$ p" means that the random variable (RV) x is distributed according to the pdf p. If k_1 and k_2 are time indices, let $k_< = \min(k_1, k_2)$ and $k_> \doteq \max(k_1, k_2)$. The Kronecker delta,function is denoted by δ_{k_1,k_2}. Superscript T denotes transpose.

2.2 Model And Signal Processing Goal

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 ($\phi_i(k)$). (The total phase signal is defined to be $\phi_i(k) = \phi_i(0) + 2\pi T \sum_{m=0}^{k-1}[f_i(m) + \nu_i(\mathrm{m})]$ where T is the sampling interval). 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 both to change values and to remain nearly constant over periods of milliseconds in duration. A random walk model is attractive because if $x(k)$ is a random walk then $\mathbf{E}[x(k)]$ is constant and $x(k) = \arg\max_{x(k+1)} p(x(k+1)|x(\mathrm{k}))$. An alternative model, an AR process with a nonzero mean μ, is not as attractive because the formant frequency will take and hold different values while only one value, the mean μ, is available in the alternative model. Generalizing the mean to be time-varying is impractical

because the time-course of its variation is not known. The dynamics of the total phase signal $\phi(k)$ are completely determined by its definition: $\phi_i(k) = \phi_i(0) + 2\pi T \sum_{m=0}^{k-1}[f_i(m) + \nu_i(m)]$ where T is the sampling interval. The measured signal, denoted by y(k), is the linear superposition of the contribution frorn each formant, specifically, $a_i(k)\cos(\phi_i(k))$, plus additive measurement noise. The complete model is therefore

$$a_i(k+1) = \alpha_{a_i} a_i(k) + q_{a_i} w_{a_i}(k) \quad (2.1)$$

$$\nu_i(k+1) = \alpha_{\nu_i} \nu_i(k) + q_{\nu_i} w_{\nu_i}(k) \quad (2.2)$$

$$f_i(k+1) = f_i(k) + q_{f_i} w_{f_i}(k) \quad (2.3)$$

$$\phi_i(k+1) = \phi_i(k) + 2\pi T f_i(k) + 2\pi T \nu_i(k) \quad (2.4)$$

$$y(k) = \sum a_i(k)\cos(\phi_i(k)) + rv(k) \quad (2.5)$$

where the process noises w_{a_i}, w_{ν_i}, and w_{f_i} and the observation noise v are all i.i.d. $\mathcal{N}(0,1)$ sequences; the covariance of the observation noise is r^2; the initial conditions are $a_i(0) \sim \mathcal{N}(0, q_{a_i}^2/(1-\alpha_{a_i}^2))$, $\nu_i(0) \sim \mathcal{N}(0, q_{\nu_i}^2/(1-\alpha_{\nu_i}^2))$, $f_i(0) \sim \mathcal{N}(m_{f_i,0}, p_{f_i,0}^2)$, and $\phi_i(0) \sim \mathcal{N}(0, p_{\phi_i,0}^2)$; 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; and ν_i are wide sense stationary random sequences. Define $\theta = (\alpha_{a_i}, q_{a_i}, \alpha_{\nu_i}, q_{\nu_i}, q_{f_i}, r, m_{f_i,0}, p_{f_i,0}, p_{\phi_i,0})^{\mathrm{T}}$.

In terms of the model, the goal of extracting amplitude and phase modulations from the observed signal corresponds to estimating $z_i(k) = (a_i(k), f_i(k), \nu_i(k), \phi_i(k))^T$ given the measurements $y(0), \ldots, y(k)$. Let $\hat{z}_i(k|l)$, a function of $y(0)$, ..., $y(l)$, be the estimate of $z_i(k)$ based on data through time 1. Let $\epsilon(k|l) \doteq \mathbf{E}[(z_i(k) - \hat{z}_i(k|l))^T(z_i(k) - \hat{z}_i(k|l))]$ be the mean square error (MSE). We define as the optimal estimator, denoted by $\hat{z}_i^*(k|l)$, that estimator which minimizes $\epsilon(k|l)$ with the result that $\hat{z}_i^*(k|l) = \mathbf{E}[z_i(k)|y(0), \ldots, y(l)]$ and the achieved MSE is $\epsilon^*(k|l) = \mathbf{E}[(z_i(k) - \hat{z}_i^*(k|l))^T(z_i(k) - \hat{z}_i^*(k|l))]$. Except in Section 2.3, in this study we are concerned with the filtering problem, for which $k = 1$. rather than prediction problems

(k > l) or smoothing problems (b < 1). Therefore, the goal of the signal processing is to compute the expectation in $\hat{z}_i^*(k|k) = \mathbf{E}[z_i(k)|y(0),\ldots,y(\mathrm{k})]$ which, however, we are only able to approximate (Section 2.5).

Note that the MSE performance criteria is not natural for all problems. In particular, since $y(k)$ is unaltered when $\phi_i(k)$ is replaced by $\phi_i(k) + l_i(k)2\pi$, where $l_i(k)$ is an integer, the MSE performance of an estimator for $\phi_i(k)$ will degrade over time as more and more errors of magnitude 2π occur in the estimate. In a frequency-modulated communication system, this is the well-known cycle-slipping phenomenon. However, the MSE performance of an estimator for $a_i(k)$, $f_i(k)$, and $\nu_i(k)$ can be free of such problems.

2.3 Cramer-Rao Bound

In order to determine whether the result of an estimation problem will have sufficient accuracy for the intended application independent of the estimation algorithm used or to compare a practical but suboptimal estimator to an absolute standard of performance, it is helpful to have a lower bound on $\epsilon^*(k|k)$. One such bound is the Cramer-Rao bound (CRB) [20, Section 2.4][21, Chapter 3][22, Section 6.4][23, Section IV.C] which we compute in this section. In Appendix B we discuss an alternative performance bound based on rate distortion theory.

There are two closely-related forms of the CRB depending on whether prior knowledge is or is not available. The scalar forms of these bounds? minus technical conditions, are

1. Cramer-Rao bound for non-random parameters [20, p. 66]: Let y be the measurement, $\boldsymbol{a}$ he the parameter, and $\hat{x}(y)$ be an unbiased estimate of x. Then

$$\mathrm{Var}[\hat{x}(y) - x] \geq \left\{ -\mathbf{E}\left[\frac{\partial^2 \ln p_{y|x}(y|"}{\partial x^2} \right] \right\}^{-1}.$$

   This bound is appropriate when x is a deterministic but unknown parameter since only the marginal probability density function $p_{y|x}$ is involved. (This is the hound that is traditionally called the CRB).

2. Cramer-Rao bound for random parameters [20, p. 72]: Let y be the measurement, x be the parameter, and $\hat{x}(y)$ be an estimate of x. Then

$$\mathbf{E}\left\{[\hat{x}(y)-x]^2\right\} \geq \left\{-\mathbf{E}\left[\frac{\partial^2 \ln p_{y,x}(y,x)}{\partial x^2}\right]\right\}^{-1}$$

This bound is appropriate when x is a random parameter since the joint probability density function $p_{y,x}$ is involved.

Because of the importance of prior knowledge in the MBDA algorithm, we use the second form of the bound in which prior knowledge is included.

Let k $= K-1$ be the time at which the CRB on $\epsilon^*(k|\mathrm{k})$ is desired. The natural approach to computing the CRB is to consider the entire trajectory of $a_i(k)$, $f_i(k)$, $\nu_i(k)$, and $\phi_i(k)$ $(i = 1,\ldots,I,\ k = 0,\ldots,K-1)$ and apply the usual CRB to this vector. The resulting bound is a CRB for the fixed-interval smoother since all of the estimates are based on the entire data, vector $y(0),\ldots,y(K-1)$. However, at time k $= K-1$, the fixed-interval smoother and the filter are identical. The difficulty with this approach is the size of the Fisher information matrix J which must be inverted: If there are $\mathbf{I} = 4$ formants and $K = 16000$ samples (1 s in the TIMIT database [24]) then J is $4IK$ x $4IK = 256000$ x 2.56000. The solution is to use the Kalman Filter (KF) to provide just the necessary block of J^{-1} [25, 26, 27].

Let $z_i(k) = (a_i(k), f_i(k), \nu_i(k), \phi_i(k))^T$, $w_i(k) = (w_{a_i}(k), w_{f_i}(k), w_{\nu_i}(k))^T$, $c_a = (1,0,0,0)^T$, and $c_\phi = (0,0,0,1)^T$. Then the model (Eqs. (2.1)-(2.5))has the form

$$z_i(k+1) = F_i z_i(k) + G_i w_i(k) \tag{2.6}$$

$$y(k) = \sum_i c_a^T z_i(k) \cos(c_\phi^T z_i(k)) + rv(k) \tag{2.7}$$

where

$$F_i = \begin{pmatrix} \alpha_{a_i} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \alpha_{\nu_i} & 0 \\ 0 & 2\pi T & 2\pi T & 1 \end{pmatrix} \quad G_i = \begin{pmatrix} \text{Yo}_i & 0 & 0 \\ 0 & q_{f_i} & 0 \\ 0 & 0 & q_{\nu_i} \\ 0 & 0 & 0 \end{pmatrix}$$

It is not possible to apply the results of Refs. [25, 26, 27] to Eqs. (2.6) and (2.7) because the covariance of the process noise in Eq. (2.6), which is $G_i G_i^T$, is not

full rank because $G_i \in \mathcal{R}^{4\times 3}$. However, the results of Ref. [27] also apply to AR processes with order greater than 1 when driven by process noise with a full rank covariance. Furthermore, it is possible to transform Eqs. *(2.6)* and *(2.7)* to such a form, while retaining the interpretation of the internal variables in Eqs. (2.6) and (2.7) as formant frequency, etc.

The first transformation is to separately sum f_i and ν_i in the system of Eqs. *(2.1)*–(2.5). Specifically, we define $\phi_{f_i}(k) \doteq \phi_{f_i}(0) + 2\pi T \sum_{m=0}^{k-1} f_i(m)$, $\phi_{\nu_i}(k) = \phi_{\nu_i}(0) + 2\pi T \sum_{m=0}^{k-1} \nu_i(m)$, and $\phi_i(k) = \phi_{f_i}(k) + \phi_{\nu_i}(k)$ and rewrite the system of Eqs. *(2.1)*–(2.5) in the form

$$
\begin{aligned}
a_i(k+1) &= \alpha_{a_i} a_i(k) + q_{a_i} w_{a_i}(k) && (2.8)\\
\nu_i(k+1) &= \alpha_{\nu_i} \nu_i(k) + q_{\nu_i} w_{\nu_i}(k) && (2.9)\\
f_i(k+1) &= f_i(k) + q_{f_i} w_{f_i}(k) && (2.10)\\
\phi_{f_i}(k+1) &= \phi_{f_i}(k) + 2\pi T f_i(k) && (2.11)\\
\phi_{\nu_i}(k+1) &= \phi_{\nu_i}(k) + 2\pi T \nu_i(k) && (2.12)\\
y(k) &= \sum_i a_i(k) \cos(\phi_{f_i}(k) + \phi_{\nu_i}(k)) + rv(k). && (2.13)
\end{aligned}
$$

The initial conditions are unchanged with tlie addition of $\phi_{f_i}(0) \sim \mathcal{N}(0, p^2_{\phi_i,0}/2)$ and $\phi_{\nu_i}(0) \sim \mathcal{N}(0, p^2_{\phi_i,0}/2)$. The second transformation is to write the pairs (f_i, ϕ_{f_i}) and (ν_i, ϕ_{ν_i}) as second-order *AR* processes, specifically,

$$
\begin{aligned}
a_i(k+1) &= \alpha_{a_i} a_i(k) + q_{a_i} w_{a_i}(k) && (2.14)\\
\phi_{f_i}(k+1) &= (1+\alpha_{f_i})\phi_{f_i}(k) - \alpha_{f_i}\phi_{f_i}(k-1) \\
&\quad + 2\pi T q_{f_i} w_{f_i}(k-1) && (2.15)\\
\phi_{\nu_i}(k+1) &= (1+\alpha_{\nu_i})\phi_{\nu_i}(k) - \alpha_{\nu_i}\phi_{\nu_i}(k-1) \\
&\quad + 2\pi T q_{\nu_i} w_{\nu_i}(k-1) && (2.16)\\
y(k) &= \sum_i a_i(k) \cos(\phi_{f_i}(k) + \phi_{\nu_i}(k)) + rv(k) \\
& && (2.17)
\end{aligned}
$$

where $\alpha_{f_i} = 1$. The initial conditions on the second-order ϕ_{f_i} process are

$$\begin{bmatrix}\phi_{f_i}(0)\\ \phi_{f_i}(1)\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\ 2\pi T m_{f_i,0}\end{bmatrix}, \begin{bmatrix}p^2_{\phi_i,0}/2 & p^2_{\phi_i,0}/2\\ p^2_{\phi_i,0}/2 & (2\pi T)^2 p^2_{f_i,0} + p^2_{\phi_i,0}/2\end{bmatrix}\right) \tag{2.18}$$

and likewise for ϕ_{ν_i}. Because the process noise has a full-rank covariance, specifically, $\mathrm{diag}(q^2_{a_i}, (2\pi T)^2 q^2_{f_i}, (2\pi T)^2 q^2_{\nu_i})$, the results of Ref. [27] can be applied to this system to compute a. lower bound on the MSE of any estimator $\hat{x}_i(k|k)$ where $x_i(k) = (a_i(k), \phi_{f_i}(k), \phi_{\nu_i}(k), a_i(k-1), \phi_{f_i}(k-1), \phi_{\nu_i}(k-1))^T$.

Because we also desire CRBs for the frequency variables f_i and ν_i, these variables must be reconstructed from ϕ_{f_i} and ϕ_{ν_i}. From Eqs. (2.11) and (2.12) we have

$$f_i(k) = (\phi_{f_i}(k+1) - \phi_{f_i}(k))/(2\pi T) \tag{2.19}$$

$$\nu_i(k) = (\phi_{\nu_i}(k+1) - \phi_{\nu_i}(k))/(2\pi T) \tag{2.20}$$

so the optimal estimates and the resulting MSE are

$$\hat{f}^*(k|k) = (\hat{\phi}^*_{f_i}(k+1|k) - \hat{\phi}^*_{f_i}(k|k))/(2\pi T) \tag{2.21}$$

$$\begin{aligned}\epsilon^*_{f_i}(k|k) = \frac{1}{(2\pi T)^2}\Big\{&\mathbf{E}[(\phi_{f_i}(k+1) - \hat{\phi}^*_{f_i}(k+1|k))^2]\\ &- 2\mathbf{E}[(\phi_{f_i}(k+1) - \hat{\phi}^*_{f_i}(k+1|k))(\phi_{f_i}(k) - \hat{\phi}^*_{f_i}(k|k))]\\ &+ \mathbf{E}[(\phi_{f_i}(k) - \hat{\phi}^*_{f_i}(k|k))^2]\Big\}\end{aligned} \tag{2.22}$$

and likewise for ν_i. Notice that both the filtering (i.e., $\hat{\phi}^*_{f_i}(k|k)$) and the one step ahead p edicting (i.e., $\phi^*_{f_i}(k+1|k)$) estimates of the phase variable ϕ_{f_i} are required in order to compute the filtering estimate of the frequency variable f_i and likewise for ν_i. Let Ξ be the CRB for $x_i(K)$ given the data $y(0), \ldots, y(K-1)$. Therefore, $\mathbf{E}[(x_i(K) - \hat{x}^*_i(K|K-1))(x_i(K) - \hat{x}^*_i(K|K-1))^T] - \Xi \geq 0$ where ≥ 0 applied to matrices means positive semi-definite. Define $e_{a_i} = (0,0,0,1,0,0)^T$, $e_{\phi_i} = (0,0,0,0,1,1)^T$, $e_{\phi_{f_i}} = (0,0,0,0,1,0)^T$, $e_{\phi_{\nu_i}} = (0,0,0,0,0,1)^T$, $e_{f_i} = (0, 1/(2\pi T), 0, 0, -1/(2\pi T), 0)^T$, $e_{\nu_i} = (0,0,1/(2\pi T), 0, 0, -1/(2\pi T))^T$. Finally, for $\alpha \in \{\mathrm{a}; ?\phi_i, \phi_{f_i}, \phi_{\nu_i}, f_i, \nu_i\}$, the CRBs are

$$\epsilon^*_\alpha(K-1|K-1) \geq e^T_\alpha \Xi e_\alpha. \tag{2.23}$$

Easy computation of Ξ using the techniques of Ref. [27] requires a third and final transformation of the system Eqs. (2.14)–(2.17). Specifically, Eq. (2.17) is replaced by

$$y(k) = (1 - \delta_{k,K}) \sum_I a_i(k) \cos(\phi_{f_i}(k) + \phi_{\nu_i}(k)) + rv(k) \tag{2.24}$$

which implies that there is no information in the measurement at time K, or equivalently, conditioning on $y(0),\ldots,y(K)$ is the same as conditioning on $y(0),\ldots,y(K-1)$.

For the computation of the CRB there is no reason to restrict attention to the system described by Eqs. (2.14), (2.15), (2.16), and (2.24) since more general systems can be considered with no additional complexity. In particular, we consider a system of the form

$$x_{k+1} = A_0 x_k + A_1 x_{k-1} + q w_k, k = 0,\ldots,K-1 \tag{2.25}$$

$$y_k = h_k(x_k) + r v_k, \quad k = -1,\ldots,K \tag{2.26}$$

where $x_k \in \mathbb{R}^n$; $y_k \in \mathbb{R}$; w_k is i.i.d. $\mathcal{N}(0, \mathbf{I},)$; v_k is i.i.d. $\mathcal{N}(0,1)$; $(x_0^T, x_{-1}^T)^T$ is $N(m^0, \Lambda^0)$; w_k, v_k, and $(x_0^T, x_{-1}^T)^T$ are independent; and $Q \doteq qq^T$ is full rank. Equivalently, the system can be written in state variable form as

$$\dot{x}_{k+1} = \begin{bmatrix} A_0 & A_1 \\ I_n & 0_n \end{bmatrix} \dot{x}_k + \begin{bmatrix} q \\ 0 \end{bmatrix} w_k, k = 0,\ldots,K-1 \tag{2.27}$$

$$y_k = h_k((I_n, 0_n)\dot{x}_k) + r v_k, \quad k = -1,\ldots,K \tag{2.28}$$

where $\dot{x}_k = (x_k^T, x_{k-1}^T)^T$. The correspondence with the system of Eqs. (2.14), (2.15), (2.16), and (2.24) is $n = 3I$, $x_k = (a_1(k), \phi_{f_1}(k), \phi_{\nu_1}(k),\ldots,a_I(k), \phi_{f_I}(k), \phi_{\nu_I}(k))^T$, $A_0 = \mathrm{diag}(A_1^{(0)} \ldots, A_I^{(0)})$, $A_i^{(0)} = \mathrm{diag}(\alpha_{a_i}, 1+\alpha_{f_i}, 1+\alpha_{\nu_i})$, $A_1 = \mathrm{diag}(A_1^{(1)},\ldots,A_I^{(1)})$, $A_i^{(1)} = \mathrm{diag}(0, -\alpha_{f_i}, -a,,,)$, $q = \mathrm{diag}(q_1,\ldots,q_I)$, $q_i = \mathrm{diag}(q_{a_i}, 2\pi T q_{f_i}, 2\pi T q_{\nu_i})$, $h_k(x) = (1 - \delta_{k,K}) \sum_{i=1}^I (d_i^T x) \cos(e_i^T x)$, $(d_i)_j = \delta_{j,3i-2}$ for $i = 1,\ldots,I$; $j = 1,\ldots,3I$, and $(e_i)_j = \delta_{j,3i-1} + \delta_{j,3i}$ for $i = 1,\ldots,\mathbf{I}$: $\mathbf{j} = 1,\ldots,3I$ where $(d_i)_j$ denotes the j th element of the vector d_i and likewise for $(e_i)_j$. The system of Eqs. (2.14), (2.15), (2.16), and (2.24) also specifies m^0 and Λ^0 but the details of the indexing are somewhat complicated and so the results are described in Appendix .4.

After extensive calculations in order to evaluate the general expressions contained in Ref. [27], we find that the Fisher information matrix for the fixed-interval smoothing problem for the system of Eqs. *(2.25)* and *(2.26)* is equal to the Fisher information matrix for the fixed-interval smoothing problem for the following linear Gaussian system:

$$x_{k+1} = A_0 x_k + A_1 x_{k-1} + q w_k, k = 0, \ldots, K-1 \tag{2.29}$$

$$\check{y}_k = C_k x_k + \check{r}\check{v}_k, \quad k = -1, \ldots, K \tag{2.30}$$

where $x_k \in R^n$; $\check{y}_k \in R^n$; w_k is i.i.d. $\mathcal{N}(0, I_n)$; $\check{v}_k$ is i.i.d. $\mathcal{N}(0, \mathbf{I}, \mathbf{!})$; $(x_0^T, x_{-1}^T)^T$ is $\mathcal{N}(m^0, \Lambda^0)$; w_k, $\check{v}_k$, and $(x_0^T, x_{-1}^T)^T$ are independent; $\check{r} \in \mathcal{R}^{n \times n}$ is defined by $\check{r} = \mathrm{diag}(r, \ldots, r)$; ancl $C_k \in \mathcal{R}^{n \times n}$ is defined *by*

$$C_k \doteq \check{r}\mathcal{H}_k^{T/2} \tag{2.31}$$

where

$$\mathcal{H}_k \doteq \frac{1}{r^2}\mathbf{E}[(\nabla_x h_k^T)(x_k)(\nabla_x h_k)(x_k)], \tag{2.32}$$

$(\nabla_x h_k)(x_k) \doteq \left(\frac{\partial h_k}{\partial (x_k)_1}, \ldots \frac{\partial h_k}{\partial (x_k)_n}\right)$, and $(x_k)_m$ denotes the m th component of the vector x_k. The system of Eqs. *(2.29)* and *(2.30)* can be written in state vector form: the state equation is Eq. *(2.27)* and the observation equation is

$$\check{y}_k = (C_k, 0_n)\dot{x}_k + \check{r}\check{v}_k, \quad k = -1, \ldots, K. \tag{2.33}$$

We now compute $\mathcal{H}_k$ for the system of Eqs. (2.14), (2.15), (2.16), and *(2.24)*. Let x be $\mathcal{N}(m, A)$. Then it is straightforward to establish the following expectations:

$$g_0(v; m, \Lambda) \doteq \frac{1}{2}\mathbf{E}[\cos(v^T x)]$$

$$= \frac{1}{2}\cos(v^T m)\exp(-\frac{1}{2}v^T \Lambda v) \tag{2.34}$$

$$g_1(s, v; m, \mathrm{A}) \doteq \frac{1}{2}\mathbf{E}[(s^T x)\sin(v^T x)]$$

$$= \frac{1}{2}\left[(s^T m)\sin(v^T m) + (s^T \Lambda v)\cos(v^T m)\right]\exp(-\frac{1}{2}v^T \Lambda v) \tag{2.35}$$

$$g_2(u, s, v; m, \Lambda) \doteq \frac{1}{2}\mathbf{E}[(u^T x)(s^T x)\cos(v^T x)]$$

$$= \frac{1}{2}\Big\{\Big[(s^T\Lambda u) + (s^T m)(u^T m) - (s^T\Lambda v)(u^T\Lambda v)\Big]\cos(v^T m) - \Big[(s^T m)(v^T\Lambda u) + (u^T m)(v^T\Lambda s)\Big]\sin(v^T m)\Big\}\exp(-\frac{1}{2}v^T\Lambda v). \quad (2.36)$$

Let m_k and Λ_k be the mean and covariance sequences for Eq. (2.219). By evaluating; $\partial h_k/\partial(x_k)_j$ and taking expectations we find that

$$\mathcal{H}_k = \frac{1-\delta_{k,K}}{r^2}\sum_{i=1}^{I}\sum_{j=1}^{I}[e_i e_j^T h_{i,j,k}^{ee} - e_i d_j^T h_{i,j,k}^{ed} - d_i e_j^T h_{i,j,k}^{de} + d_i d_j^T h_{i,j,k}^{dd}] \quad (2.37)$$

where

$$h_{i,j,k}^{ee} \doteq \mathbf{E}\left[(d_i^T x_k)\sin(e_i^T x_k)(d_j^T x_k)\sin(e_j^T x_k)\right] = g_2(d_i, d_j, e_i - e_j; m_k, \Lambda_k) - g_2(d_i, d_j, e_i + e_j; m_k, \Lambda_k) \quad (2.38)$$

$$h_{i,j,k}^{ed} \doteq \mathbf{E}\left[(d_i^T x_k)\sin(e_i^T x_k)\cos(e_j^T x_k)\right] = g_1(d_i, e_i + e_j; m_k, \Lambda_k) + g_1(d_i, e_i - e_j; m_k, \Lambda_k) \quad (2.39)$$

$$h_{i,j,k}^{de} \doteq \mathbf{E}\left[\cos(e_i^T x_k)(d_j^T x_k)\sin(e_j^T x_k)\right] = g_1(d_j, e_i + e_j; m_k, \Lambda_k) - g_1(d_j, e_i - e_j; m_k, \Lambda_k) \quad (2.40)$$

$$h_{i,j,k}^{dd} \doteq \mathbf{E}\left[\cos(e_i^T x_k)\cos(e_j^T x_k)\right] = g_0(e_i - e_j; m_k, \Lambda_k) + g_0(e_i + e_j; m_k, \Lambda_k). \quad (2.41)$$

The algorithm for computing the CRBs is

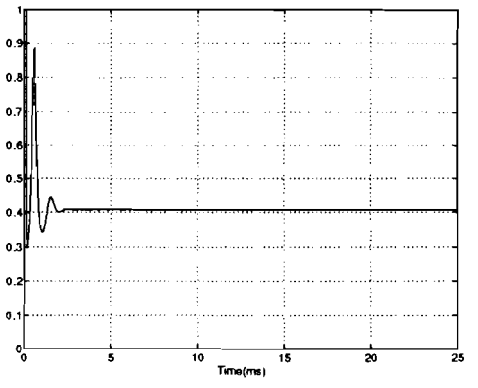
1. Fix K.
2. Compute m_k and Λ_k for $k = -1, \ldots, K$ **by** using Eq. (2.27) and standard linear system formulae.
3. Compute C_k for $k = -1, \ldots, K$ by using Eqs. (2.31) and (2.37)–(2.41).
4. Apply standard Kalman filtering formulae to the system of Eqs. ('2.27) and (2.33) to derive the MSE for time $k = K$. This $6I$ x $6I$ matrix is the CRB for

the filtering problem a.t time $k = \mathbf{Ii'}$ for the nonlinear system of Eqs. (2.27) and (2.28).

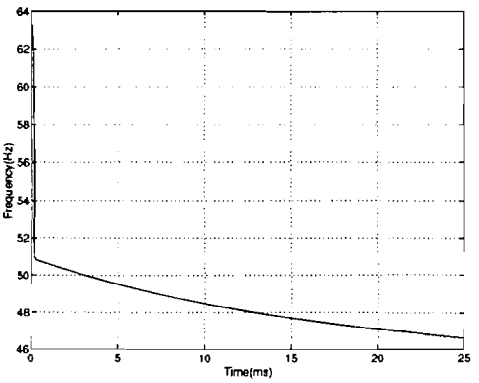
5. From the matrix resulting from Step 4, use Eq. (2.23) to compute CRBs for the filtering problem for $f_i(K)$, etc.

When K is changed, most of this work does not need to be redone because $\mathcal{H}_k$ is independent of K except at $k = \mathbf{Ii'}$.

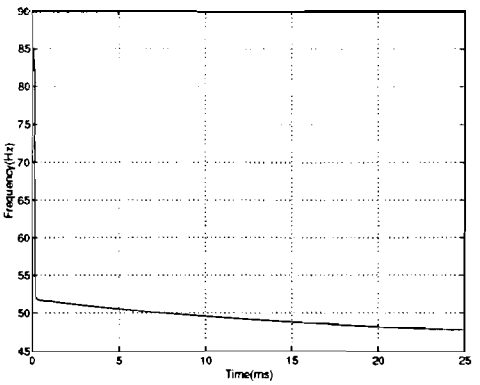
For parameters typical of the speech models used in Sections 3.2, 3.3, and 3.4, the CRBs as a function of time are shown in Figure 2.1. The model has 1 formant and parameters T = 1/16000 s, $\mathbf{a}_{\nu_1} = \alpha_{\nu_1} = .99$, $q_{a_1} = 50$, $q_{\nu_1} = 12$, $q_{f_1} = 2$, $m_{f_1,0} = 500$ Hz, $r = \sqrt{1/12}$, $P^2_{\phi_1,0} = .01$ (essentially zero), $P^2_{f_1,0} = 4000$. The oscillations in the CRB standard deviation for a_1 occur at about twice the formant frequency of 500 Hz and are due to the fact that when the cosine of $a_1(k)\cos(\phi_1(k))$ passes through 0 there is no information in $y(k)$ about $a_1(k)$ while when the cosine passes through ± 1 there is maximal information. Relative to the precision needed in the speech application and relative to the *a priori* (i.e., no measurements) standard deviations of the various signals, these CRB standard deviations are small: (1) The CRB standard deviations for the f_1 and ν_1 frequencies are less than 10% of the 500 Hz formant frequency. (2) The CRB standard deviation for ν_1 is also much smaller than the steady-state *a priori* standard deviation of 85.1 Hz. **(3)** For f_1 there is no steady state and, in fact, the *a priori* standard deviation grows as $\sqrt{k}q_{f_1}$, so the CRB standard deviation is dramatically lower. (4) The parameters used in this example are appropriate for the TIMIT database [24] where typical large signal values are 10^3 and therefore the CRB standard deviation for a_1 is less than .1%of typical large signal values. *(5)* Furthermore, the CRB standard deviation for a_1 is much less than the steady-state a *priori* standard deviation of 354. Because the CRB bounds are lower than the estimation standard deviations required by the speech application, it is worthwhile to design nonlinear filters based on this statistical model and, in Section 2.5, we describe nonlinear filters which, as shown in Sections 3.1–3.4.



(a)



(b)



(c)

Fig. 2.1. Cramer-Rao bounds for (a) $a_1(k)$, (h) $f_1(k)$, and (c) ν_1(k). The standard deviation, rather than variance, is shown. The bound for estimation of $f_1(k)$ ($\nu_1(k)$) decreases to 45.151 (46.5) at 62.5 ms (62.5 ms).

achieve good performance.

2.4 System Identification

In order to use the model of Eqs. (2.1)–(2.5), it is necessasy to determine the parameter vector θ. In view of the importance of spectral ideas in speech processing, we choose 8 by fixing the center frequency, bandwidth, and power of each formant. Therefore we need to compute $S_y(\Omega)$, the spectrum of the model (Eqs. (2.1)–(2.5)), as a function of the parameters 8, which is the subject of the following two paragraphs.

Let $a(k)$ and $\phi(k)$ be random sequences where ϕ is Gaussian. Let ψ he a RV that is distributed uniformly on [−n,n]. Let a, ϕ, and ψ be independent. Define $y(k) = a(k)\cos(\phi(k) + \psi)$. It follows that $m_y(k) = 0$ and

$$\begin{aligned} R_y(k_1, k_2) &= P_y(k_1, k_2) \\ &= R_a(k_1, k_2)\frac{\cos(m_\phi(k_1) - m_\phi(k_2))}{2} \exp\left(-\frac{1}{2}Q_\phi(k_1, k_2)\right) \end{aligned} \tag{2.42}$$

where $Q_\phi(k_1, k_2) \doteq P_\phi(k_1, k_1) - 2P_\phi(k_1, k_2) + P_\phi(k_2, k_2)$. More generally, if $y(k) = \sum_i y_i(k) + rv(k)$ where $y_i(k) = a_i(k)\cos(\phi_i(k)+\psi_i)$; a_i, ϕ_i, ψ_i and v are independent; and, tor each i, the quantities a_i, ϕ_i, and ψ_i are as above, then it follows that $m_y(k) = rm_v(k)$ and $P_y(k_1, k_2) = \sum_i P_{y_i}(k_1, k_2) + r^2 P_v(k_1, k_2)$ where P_{y_i} is given by Eq. (2.42). Since $Q_\phi(k_1, k_2)$ can alternatively be expressed as $Q_\phi(k_1, k_2) = \mathbf{E}[\{[\phi(k_1) - m_\phi(k_1)] - [\phi(k_2) - m_\phi(k_2)]\}^2]$ it follows that $Q_\phi(k_1, k_2) \geq 0$.

For the system of Eqs. (2.1)–(2.5) with $|\alpha_{a_i}| < 1$ and $|\alpha_{\nu_i}| < 1$, it follows that $R_{a_i}(k_1, k_2) = (q_{a_i}^2/(1 - \alpha_{a_i}^2))\alpha_{a_i}^{|k_2 - k_1|}$, $m_{\phi_i}(k) = 2\pi T m_{f_i,0} k$, $R_v(k_1, k_2) = \delta_{k_1,k_2}$, and

$$\begin{aligned} Q_{\phi_i}(k_1, k_2) = (2\pi T)^2 \Bigg\{ & p_{f_i,0}^2 (k_2 - k_1)^2 \\ & + \frac{q_{f_i}^2 |k_2 - k_1|(k_>(k_> - 1) - k_<(k_< - 1))}{2} \\ & - \frac{q_{f_i}^2 (|k_2 - k_1| + 1)|k_2 - k_1|(|k_2 - k_1| - 1)}{6} \\ & + \frac{q_{\nu_i}^2}{1 - \alpha_{\nu_i}^2}\left[\frac{(1 + \alpha_{\nu_i})|k_2 - k_1|}{1 - \alpha_{\nu_i}} - \frac{2\alpha_{\nu_i}(1 - \alpha_{\nu_i}^{|k_2 - k_1|})}{(1 - \alpha_{\nu_i})^2}\right]\Bigg\} \end{aligned}$$

Using these results in Eq. (2.42) provides the necessary $R_y(k_1, k_2)$ for the system of Eqs. (2.1)–(2.5).

Because of the second term in the braces, $Q_{\phi_i}(k_1, k_2)$ is not a function of $k_1 - k_2$ and therefore $R_y(k_1, k_2)$ is not wide-sense stationary. This reflects the fact that the speech signal itself is not stationary except over short intervals of time. Therefore. for choosing parameters, we set $q_{f_i} = 0$ and then, for use in the nonlinear filter, we reset q_{f_i} to the maximum desired change in the i th formant frequency per sample.

We take $p_{f_i,0} = 0$ and $p_{\phi_i,0} = 0$. The value of r is set from *a priori* knowledge of the observation noise process. For signals from the essentially noise-free TIMIT database [24], the observation noise is just the quantization noise which, under a uniform $\pm 1/2$-bit model, has standard deviation $r = \sqrt{1/12}$.

For stability we require that $|a_{i,1}| < 1$ and $|\alpha_{\nu_i}| < 1$ and to minimize the bandwidth we desire $a_{i,1} \approx 1$ and $\alpha_{\nu_i} \approx 1$. We have taken $\alpha_{a_i} = a_{i,1} = .99$ where the equality $\alpha_{a_i} = a_{i,1}$ is motivated by the error bounds of Refs. [5, 4].

It remains only to pick $m_{f_i,0}$, q_{a_i}, and q_{ν_i}. The center frequency of the formant is $m_{f_i,0}$. The only effect of q_{a_i} is to scale R_{a_i} and so it does not effect the bandwidth of the formant. Therefore, we use q_{ν_i} to fix the bandwidth of the formant, according to the plot of Figure 2.2. The resulting S_{y_i} are shown in Figure 2.4 for a variety of choices of q_{ν_i}. Once q_{ν_i} is fixed, we use q_{a_i} to fix the power of the formant: use the plot of Figure 2.3 to determine the power that would be present if q_{a_i} equaled 1 and set q_{a_i} to scale this to the desired value. Finally, for use in the nonlinear filter of Set-tion 2.5, the value of q_{f_i} is reset to the maximum desired change in the i th formant frequency per sample. In all calculations we have computed S_{y_i} from R_{y_i} by computing the DFT of the sequence $R_{y_i}(0, k)$ for $k = -4096, \ldots, \$4096$.

2.5 Nonlinear Filters

If $a_i(k)$ was constant then Eqs. (2.1)–(3.5) describe a frequency modulated communication system, the Extended Kalman filter (EKF) [28, Section 8.2] is essentially a phase-locked loop (PLL), and the PLL is an excellent estimator. Therefore, we

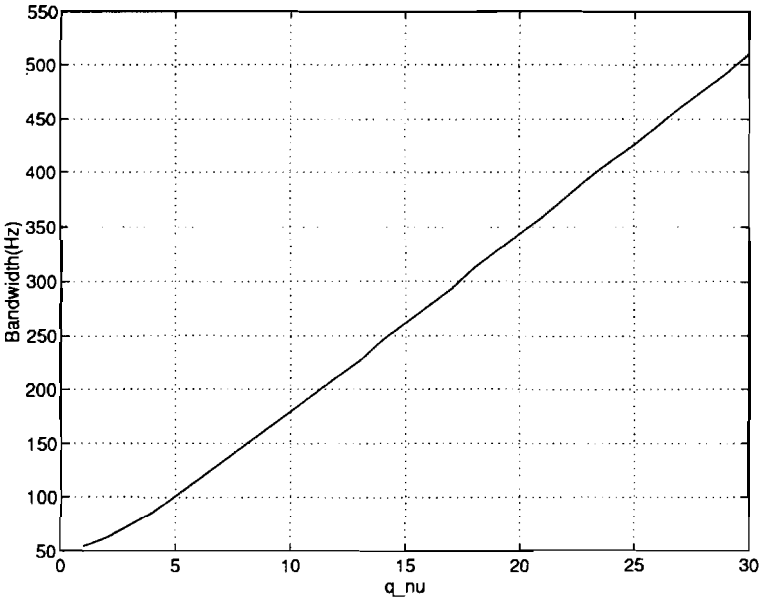


Fig. 2.2. Half-power (3 dB) bandwidth of S_{y_i} as a function of q_{ν_i}. The parameters are $\alpha_{a_i} = .99$, $q_{a_i} = 1$, $\alpha_{\nu_i} = .99$, $q_{f_i} = 0$, $r = 0$, $p_{f_i,0} = 0$, $p_{\phi_i,0} = 0$, and T = 1/16000 s. $m_{f_i,0}$ does not affect the bandwidth.

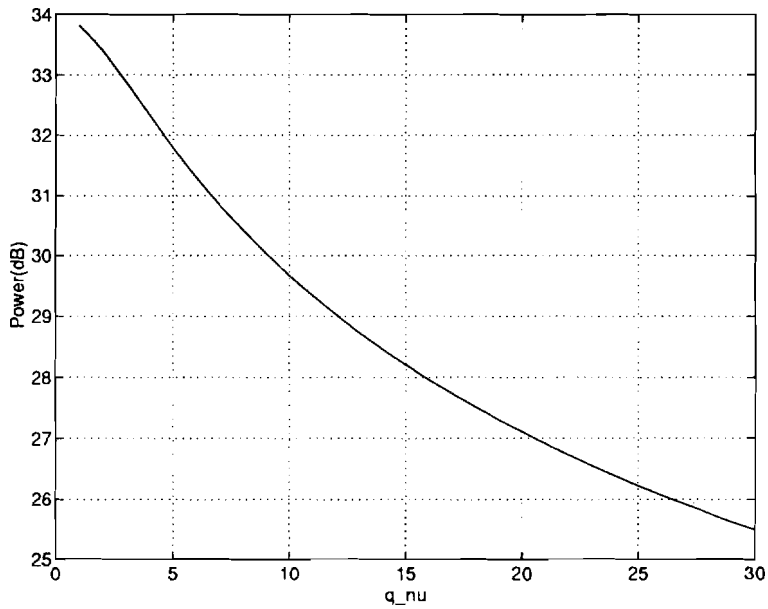


Fig. 2.3. Peak power of S_{y_i} as a function of q_{ν_i}. The parameters are $a_{,,} = .99$, $q_{a_i} = 1$, $\alpha_{\nu_i} = .99$, $q_{f_i} = 0$, $r = 0$, $p_{f_i,0} = 0$, $p_{\phi_i,0} = 0$, and T = 1/16000 s, $m_{f_i,0}$ does not affect the peak power.

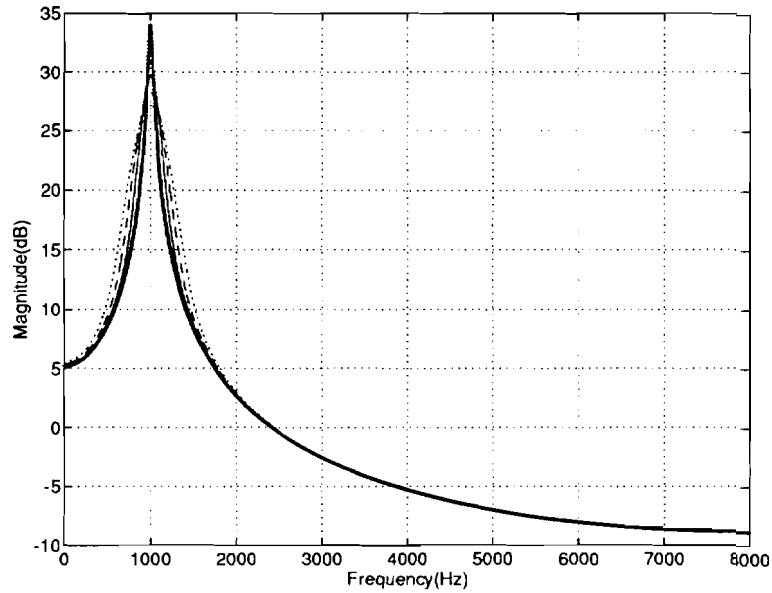


Fig. 2.4. Example S_{y_i} curves. The parameters are $a_{,,} = .99$, $q_{a_i} = 1$, $\alpha_{\nu_i} = .99$. $q_{\nu_i} = .1, 1, 10, 15, 20$, $q_{f_i} = 0$, $r = 0$, $m_{f_i,0} = 1000$ Hz, $p_{f_i,0} = 0$, $p_{\phi_i,0} = 0$, and T = 1/16000 s, The peaks are broader as q_{ν_i} in reases.

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, me will not indicate the conditioning which is always $k|k$) by using the EKF for this more corn-plicated model. The computational requirements are minimal: the state equation is already linear, the one-step state transition matrix (denoted by $\mathbf{F}$) is block diagonal (1 block per formant) and each block is sparse so multiplication by $\mathbf{F}$ 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 $\hat{y}(k) = \sum_i \hat{a}_i(k) \cos(\hat{\phi}_i(k))$.

3. APPLICATIONS OF MBDA

In this chapter, we apply the statistical model and the nonlinear estimator discussed in the previous chapter to some synthetic and real speech problems. We consider three synthetic examples (Section 3.1), decomposition of speech into AM and FM signals (Section 3.2), two formant tracking problems: transitions to stops (Section 3.3) and tracking formants through a sentence (Section 3.4), and application to unvoiced speech (Section 3.5).

3.1 Application To Synthetic Examples

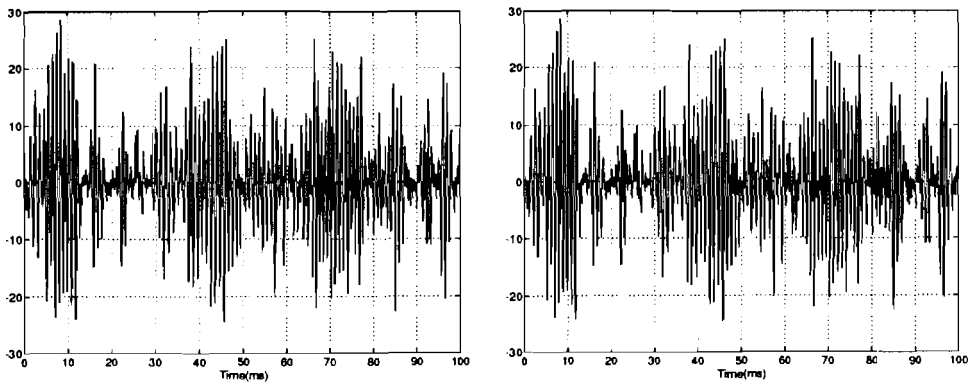
In the first example we demonstrate the effectiveness of the EKF by successfully processing a synthetic signal that is a realization of the model Eqs. (2.1)–(2.5). The model has 1 formant with initial condition $m_{f_1,0} = 1000$ Hz. The other parameters are $T = 1/16000$ s, $a_{,,} = a_{,,} = .99$, $q_{a_1} = 2$, $q_{f_1} = 2$, $y_{,,} = 15$, and $r = \sqrt{1/12}$.

In Figures 3.1 and 3.2 we show the original and reconstructed signals in the time and frequency domains respectively. In Figure 3.3 we show the true and estimated trajectories for $a_1(k)$, $\nu_1(k)$, $f_1(k)$, and $\phi_1(k)$ over an interval of 100 ms.

In the second example we apply EKF to a chirp signal patterned after [4, Figure 2]

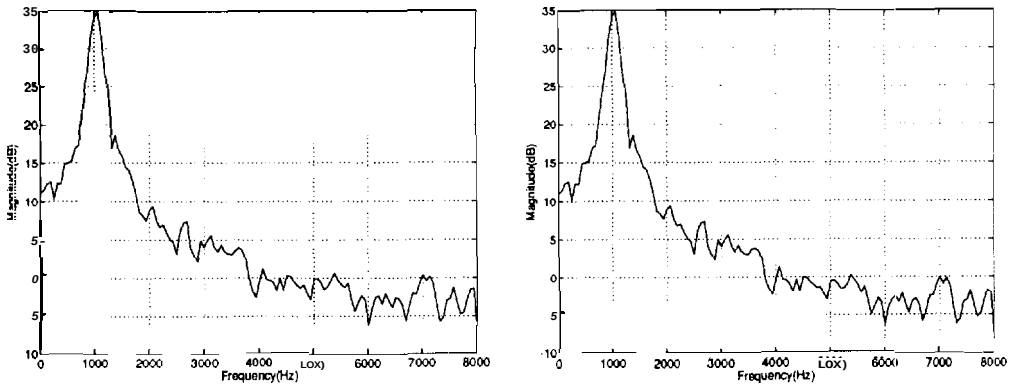
$$y(k) = \cos(2\pi f_a kT)\cos(2\pi(f_m + f_c kT)kT), \tag{3.1}$$

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). We use the model of Eqs. (2.1)–(2.3)with 1 formant. The parameters are $a_{,,} = a_{,,} = .99$, $q_{a_1} = 0.1$, $q_{f_1} = 3$, $q_{\nu_1} = 0.1$, $r = \sqrt{1/12}$



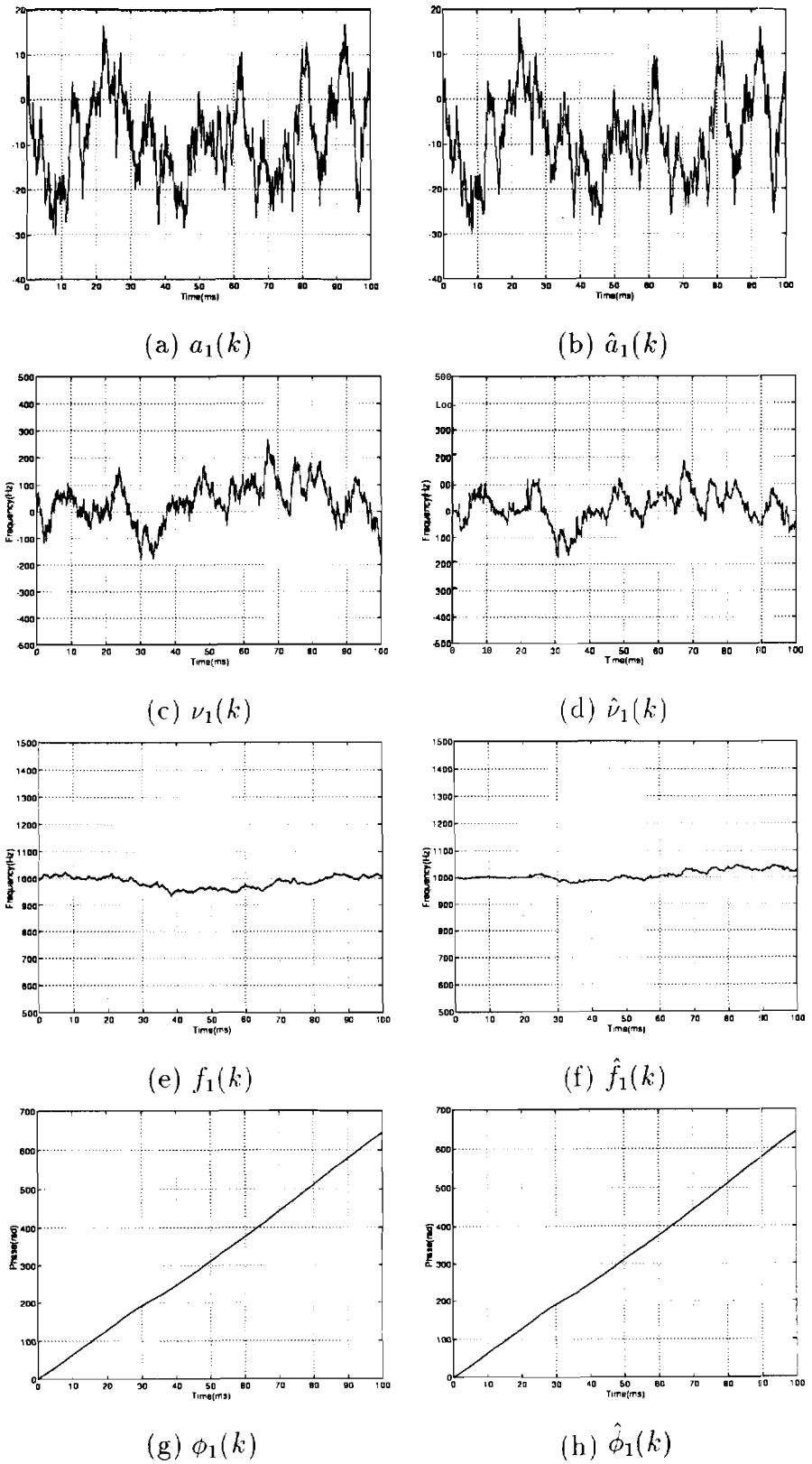
(a) Original (b) Reconstructed

Fig. 3.1. The original and reconstructed synthetic signals in the time domain.



(a) Original (b) Reconstructed

Fig. 3.2. The synthetic signals in the frequency domain: Power spectral density (Welch method with a 256 point FFT and 50% overlap) of the signals in Figure 3.1.



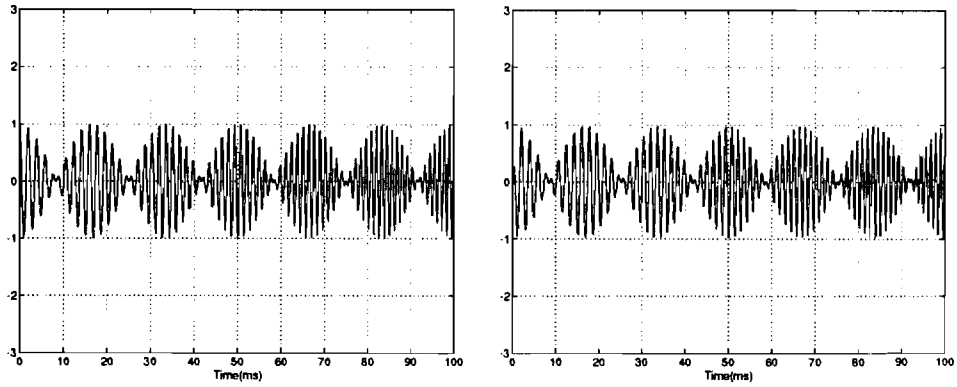
(a) $a_1(k)$ (b) $\hat{a}_1(k)$

(c) $\nu_1(k)$ (d) $\hat{\nu}_1(k)$

(e) $f_1(k)$ (f) $\hat{f}_1(k)$

(g) $\phi_1(k)$ (h) $\hat{\phi}_1(k)$

Fig. 3.3. True and estimated trajectories for the synthetic signal.



(a) $y(k)$ (b) $\hat{y}(k)$

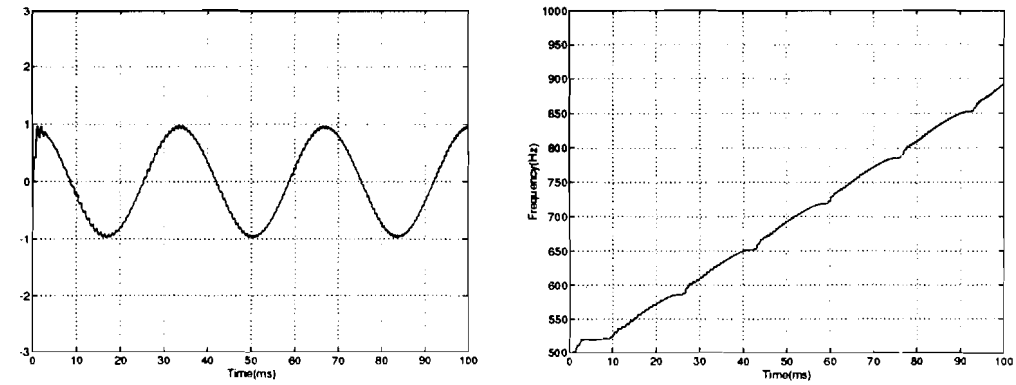
Fig. **3.4.** The original ($y(k)$) and reconstructed ($\hat{y}(k)$) one-chirp synthetic signals.

and $m_{f_1,0} = f_m$. In Figure 3.4 we show the original ancl reconstructed signals. In Figure **3.5** we show the 4 EKF outputs. Note the instantaneous frequency in the signal is $f_m + 2f_c kT$ while the instantaneous frequency in our model is $f(k) + \nu(k)$. The results, shown in Figure 3.5, 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 $a(k)$.

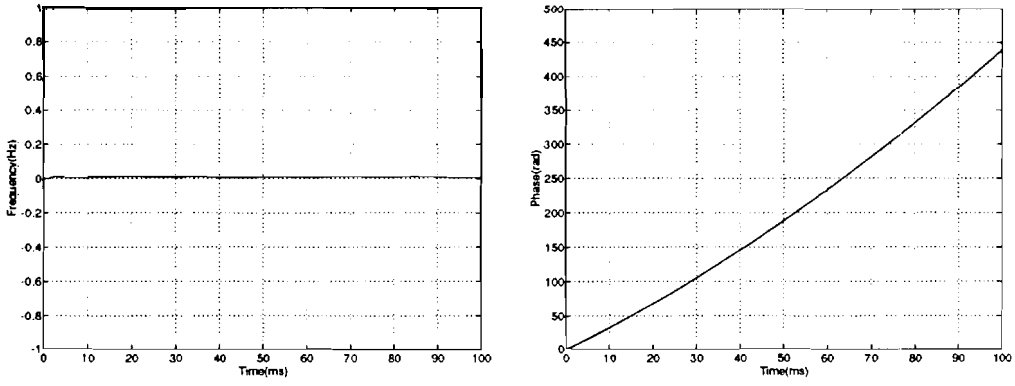
In the third example, we show the results of applying the model (Eqs. (2.1)–(2.5)) and EKF to a double chirp signal [29] patterned after the single chirp signal of Maragos, Kaiser, and Quatieri. The signal is

$$y(\mathrm{k}) = \cos(2\pi f_a kT)\cos(2\pi(f_{m1}+f_c kT)kT)+0.2\cos(2\pi f_a kT)\cos(2\pi(f_{m2}-f_c kT)kT)$$

where T = 1/16000 s, f_a = 30 Hz, f_{m1} = 200 Hz, f_{m2} = 2000 Hz, and f_c = 6000 Hz/s. The EKF has parameters $a_{,,} = a_{,,} = .99$, $a_{,,} = a_{,,} = .99$, $q_{a_1} = \sqrt{.001}$, $q_{a_2} = \sqrt{.001 \times .04}$, $q_{f_1} = q_{f_2} = \sqrt{12}$, $q_{\nu_1} = q_{\nu_2} = \sqrt{.01}$, $r = 1.0 \times 10^{-7}$, $m_{f_1,0} = f_{m1}$, $m_{f_2,0} = f_{m2}$, and all the other initial condition means and all the initial condition variances equal to zero. Ideal performance of the EKF in this example would lead to $\hat{a}_1(k) = \cos(2\pi f_a kT)$, $\hat{f}_1(k) = f_{m1} + 2f_c kT$, $\hat{a}_2(k) = 0.2\cos(2\pi f_a kT)$, $\hat{f}_2(k) = f_{m2} - 2f_c kT$, and $\hat{\nu}_1(k) = \hat{\nu}_2(k) = 0$. The actual results, shown in Figures 3.6–3.8,

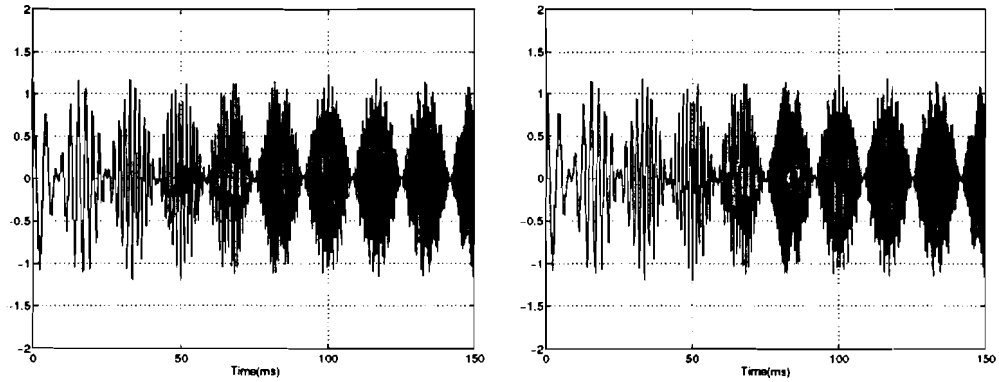


(a) $\hat{a}_1(k)$ (b) $\hat{f}_1(k)$



(c) $\hat{\nu}_1(k)$ (d) $\hat{\phi}_1(k)$

Fig. 3.5. EKF estimates for the one-chirp synthetic signal.



$y(k)$ $\hat{y}(k)$

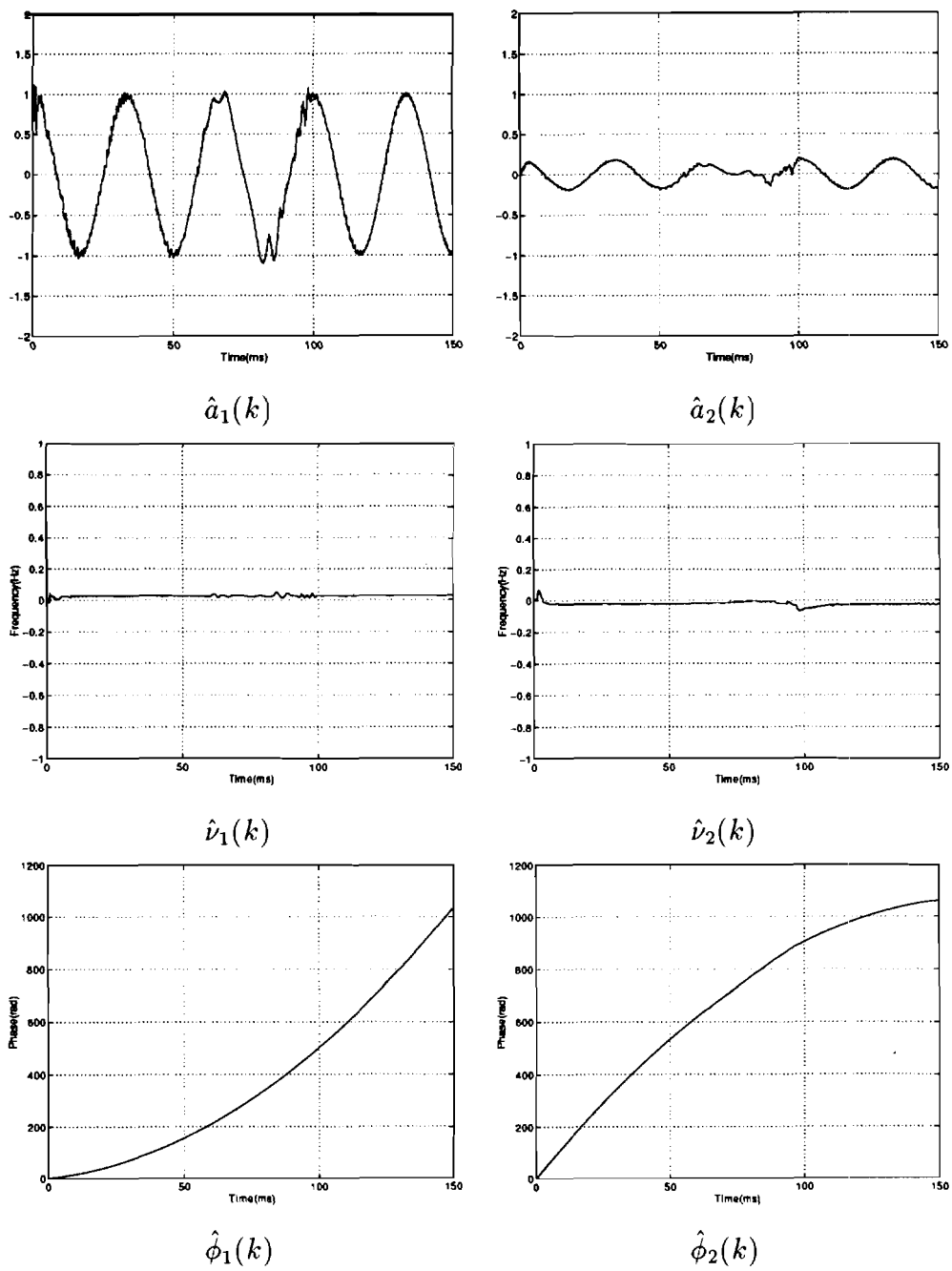
Fig. 3.6. The original ($y(k)$) and reconstructed ($\hat{y}(k)$) two-chirp synthetic signals.

are quite good.

3.2 Application To Speech

In this section we show the results of applying the moclel (Eqs. (2.1)–(2.5)) ancl EKF to a speech signal. The signal is the phoneme /ee/ of the word "m/ee/ting" from the TIMIT database [24, dr2/mdbb0/sx295]. The model has 4 formants with initial contlitions $m_{f_i,0}$ of 390, 2200, 2800, and 3600 Hz for i = 1. 2, 3, and 4 respectively. For all 4 formants, $\alpha_{a_i} = a_{\nu_i} = .99$, $q_{f_i} = \sqrt{8}$, $p_{f_i,0} = 0$, and $p_{\phi_i,0} = 0$. The values of q_{a_i} and q_{ν_i} vary from formant to formant: $q_{a_i} = 158, 20, 11, 7$; and $q_{\nu_i} = 14, 14, 21, 21$ for $i = 1, 2, 3, 4$ respectively. Finally, $r = \sqrt{1/12}$.

In Figures 3.9 and 3.10 we show the original and reconstructed speech in the time and frequency domains respectively. The only visible differences are in the frequency domain in two frequency bands: the band near 1500 Hz, where the signal strength is down by 40 dB from nearby formant peaks, and in the frequencies greater than 7 kHz, where the highest frequency formant in the moclel is at much lower frequency, specifically, at 3.6 kHz, and the signal strength is down by 40 dB. In Figures 3.11, 3.12 and 3.13 we show the estimates from the EKF. If smoother estimates are desired, they coulcl be achieved either by post-processing [4] or by

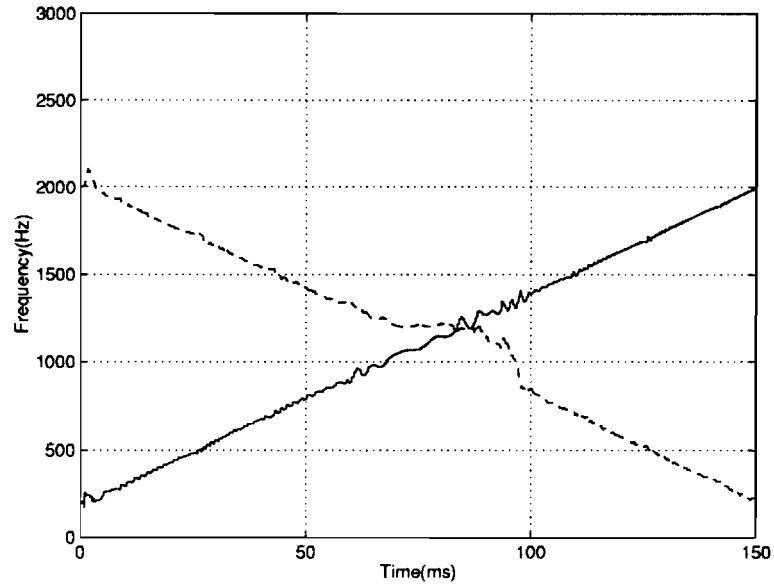


$\hat{a}_1(k)$ $\hat{a}_2(k)$

$\hat{\nu}_1(k)$ $\hat{\nu}_2(k)$

$\hat{\phi}_1(k)$ $\hat{\phi}_2(k)$

Fig. 3.7. EKF estimates for the two-chirp synthetic signal..



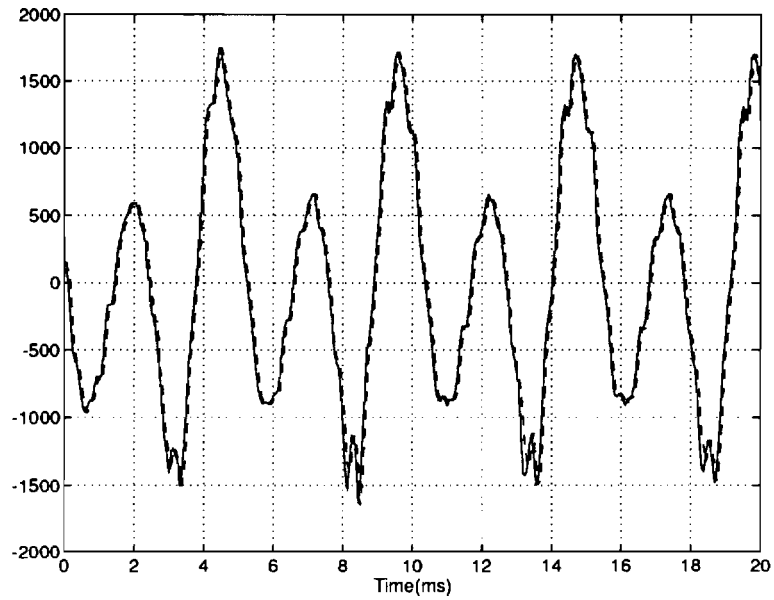
$\hat{f}_1$(k) (solid curve) and $\hat{f}_2(k)$ (dashed curve)

Fig. 3.8. EKF estimates of the frequencies for the two-chirp synthetic signal.

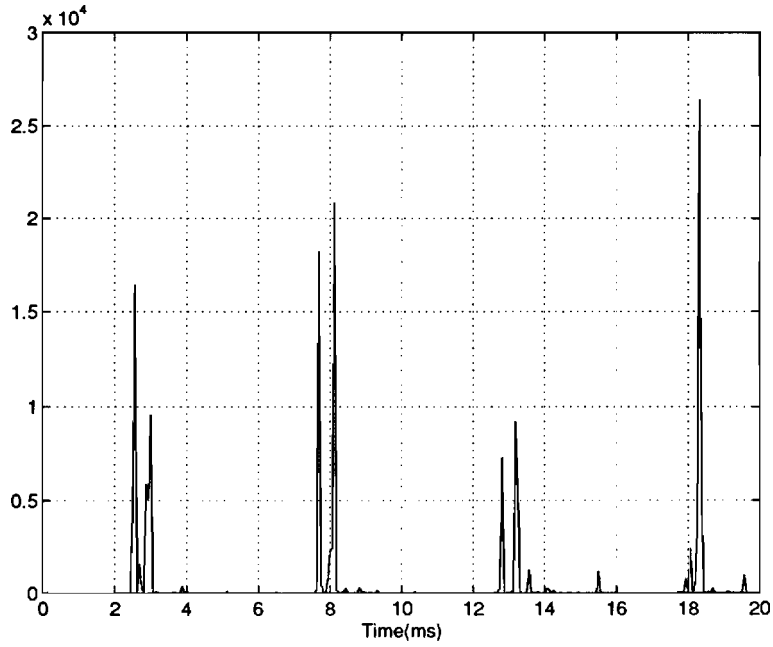
increasing α_{a_i} and $a_{,,}$ and decreasing q_{a_i} and q_{ν_i}, which changes the moclel so that the model describes signals of the same power but longer correlation time, and increasing r, which relaxes the penalty on non-exact decomposition (i.e., y(k) $\neq \hat{y}(k)$). As in Ref. [4], the major observation is the pitch-synchronous pulse structure of the Kaiser-Teager amplitude ($\hat{a}_i(k)$) and frequency ($\hat{\nu}_i(k)$) signals. Because the MBDA algorithm estimates $a_i(k)$ while the DESA-1 algorithm [4] based on Teager's energy operator estimates $|a_i(k)|$, the pulses in the estimate of $a_i(k)$ in Figure 3.11 and 3.12 are of alternating sign while the pulses in the estimate of $|a_i(k)|$ from DESA-1 (e.g., Figures 5–7 of Ref. [4]) are all positive-going.

3.3 Formant Tracking: Transitions To Stops

One application of formant tracking is to aid in the classification of stops [9, 10, 11]. In this section we apply the rnodel (Eqs. (2.1)–(2.5))and EKF to this problem and show the results for the phoneme /u/ of the word "c/u/ps" frorn the TIMIT



(a)



(b)

Fig. **3.9.** The phoneme /ee/ of the word m/ee/ting in the time domain.
(a) Original (solid curve) and reconstructed (dashed curve) speech signals.
(b) Square error, i.e., $[y(k) - \hat{y}(k)]^2$.

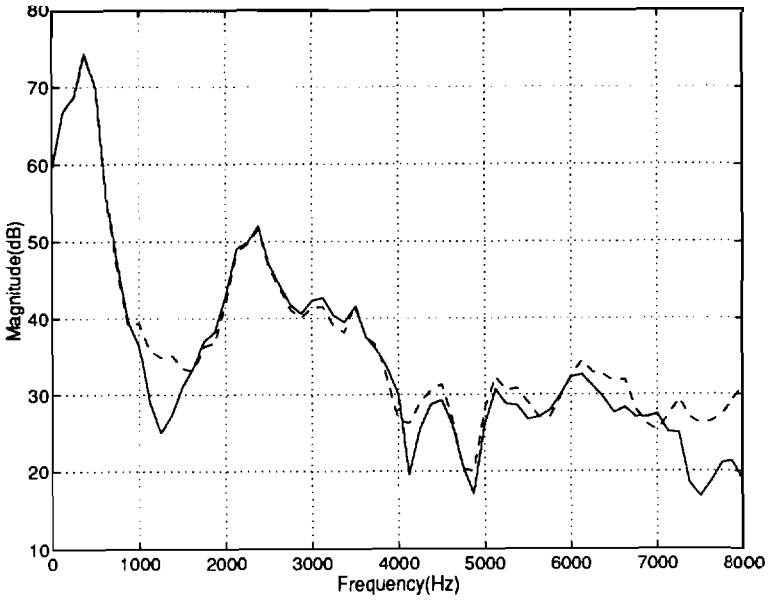
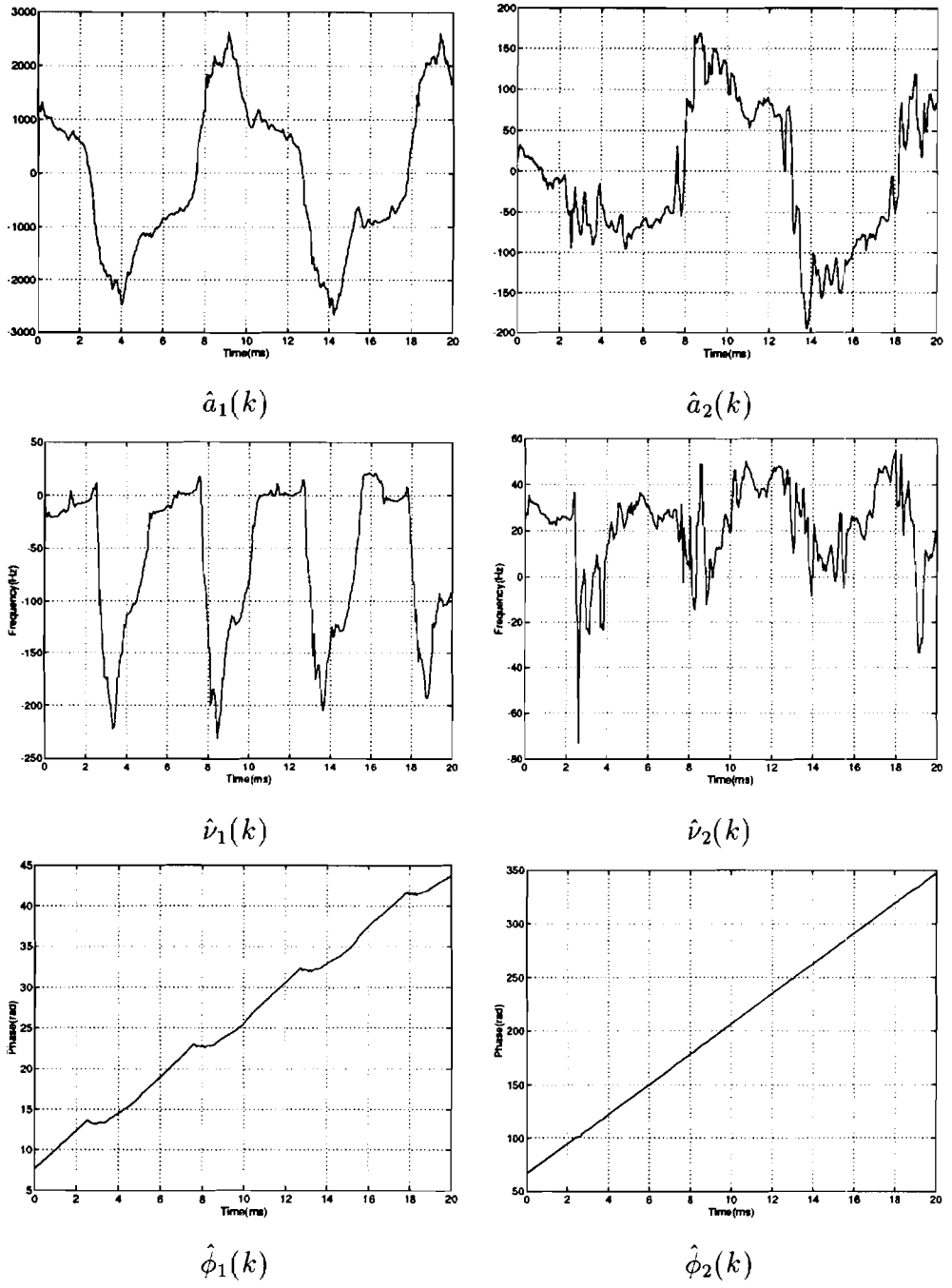


Fig. 3.10. The phoneme /ee/ of the word m/ee/ting in the frequency domain: Power spectral density (Welch method with a 128 point FFT and 50% overlap) of the signals in Figure 3.9. Original: solid curve. Reconstructed: dashed curve.

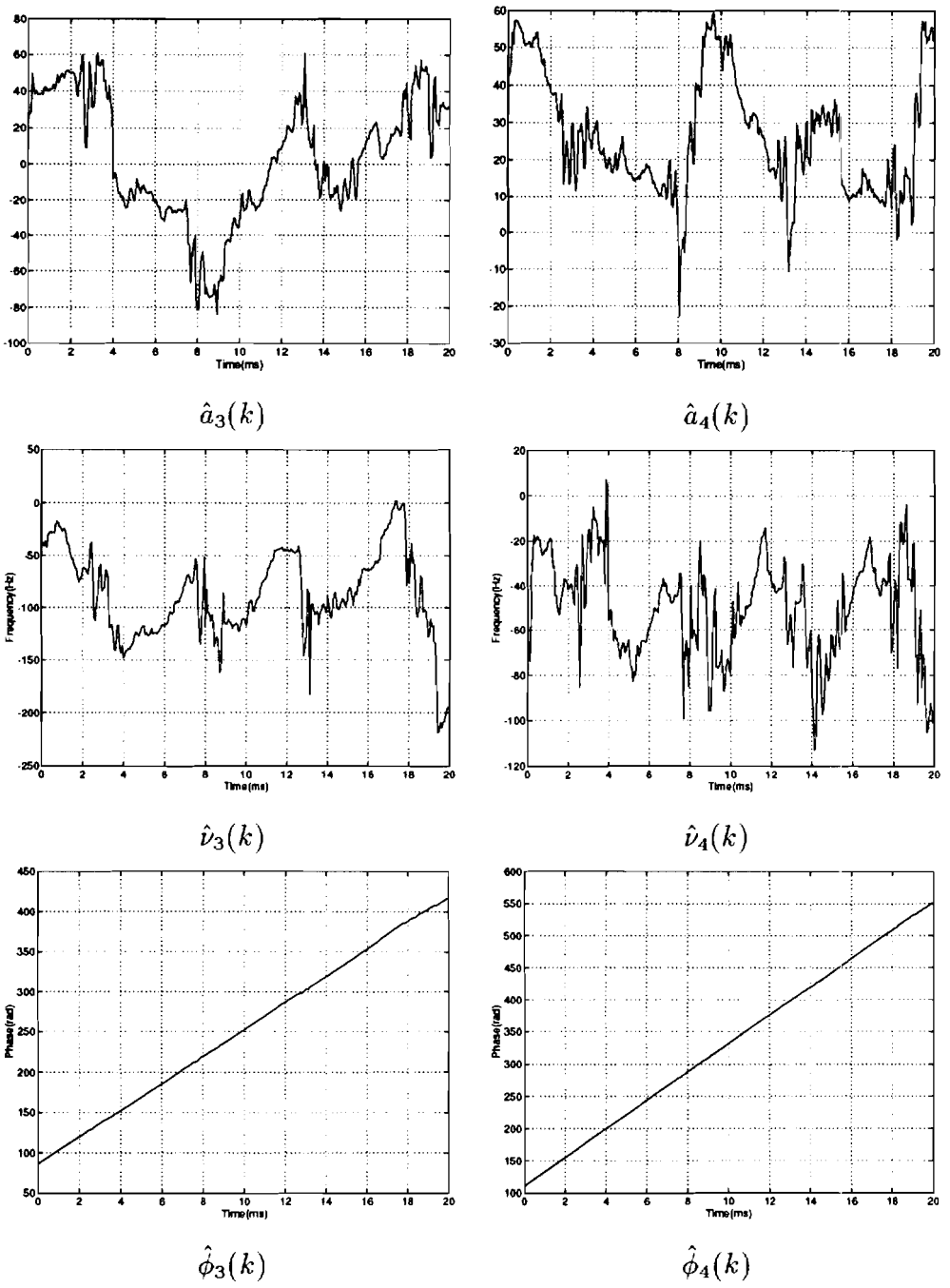


$\hat{a}_1(k)$ $\hat{a}_2(k)$

$\hat{\nu}_1(k)$ $\hat{\nu}_2(k)$

$\hat{\phi}_1(k)$ $\hat{\phi}_2(k)$

Fig. 3.11. EKF estimates for the phoneme /ee/ of the word m/ee/ting: $i = 1, 2$.



$\hat{a}_3(k)$ $\hat{a}_4(k)$

$\hat{\nu}_3(k)$ $\hat{\nu}_4(k)$

$\hat{\phi}_3(k)$ $\hat{\phi}_4(k)$

Fig. 3.12. EKF estimates for the phoneme /ee/ of the word m/ee/ting: $i = 3, 4$.

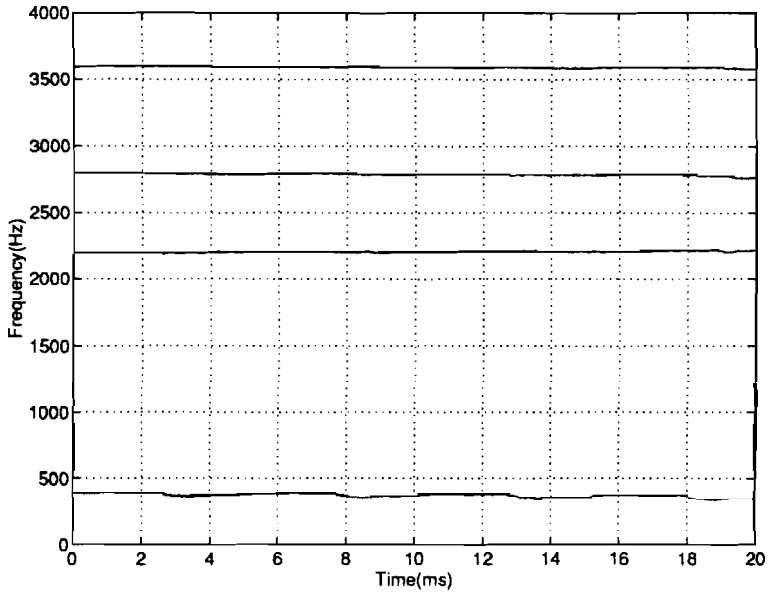


Fig. 3.13. EKF estimates for the phoneme /ee/ of the word m/ee/ting: the four formant signals $\hat{f}_1$(k), $\hat{f}_2(k)$, $\hat{f}_3(k)$, and $\hat{f}_4(k)$ (from bottom to top).

database [24, dr3/mctw0/si743]. The model has 4 formants with initial conditions $m_{f_i,0}$ of 670, 1100, 2400, and 4000 Hz for i = 1, 2, 3, and 4 respectively. For all 4 formants, $\boldsymbol{a}_{,,} = \alpha_{\nu_i} = .99$, $q_{f_i} = 2$, $p_{f_i,0} = 0$, and $p_{\phi_i,0} = 0$. The values of q_{a_i} and q_{ν_i} vary from formant to formant: $q_{a_i} = 50, 14, 3, 1$; and $q_{\nu_i} = 17.17, 15, 20$ for $i = 1, 2, 3, 4$ respectively. Finally, $r = \sqrt{1/12}$. In Figure 3.14 we show the first 2 formants at the transition. The trend is for the frequencies of both the first and second formants to decrease. A similar trend of decreasing frequencies for the lowest two formants during the transition to the unvoiced stop consonant /p/ was found by Nathan, Lee, aad Silverman [9, Figure 7a]. In the second formant, the same trend was found by Foote, Mashao, and Silverman [11, Figure 3] using the DESA-1 algorithm [4] based on the Kaiser-Teager energy operator.

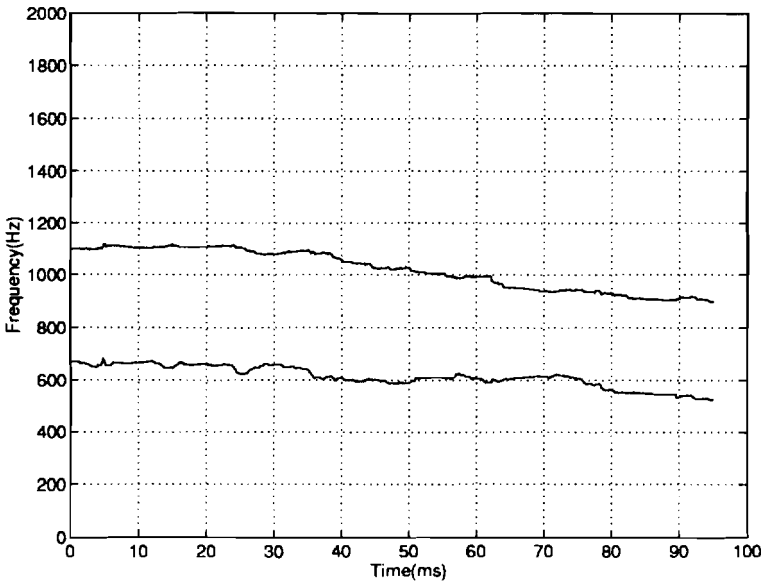
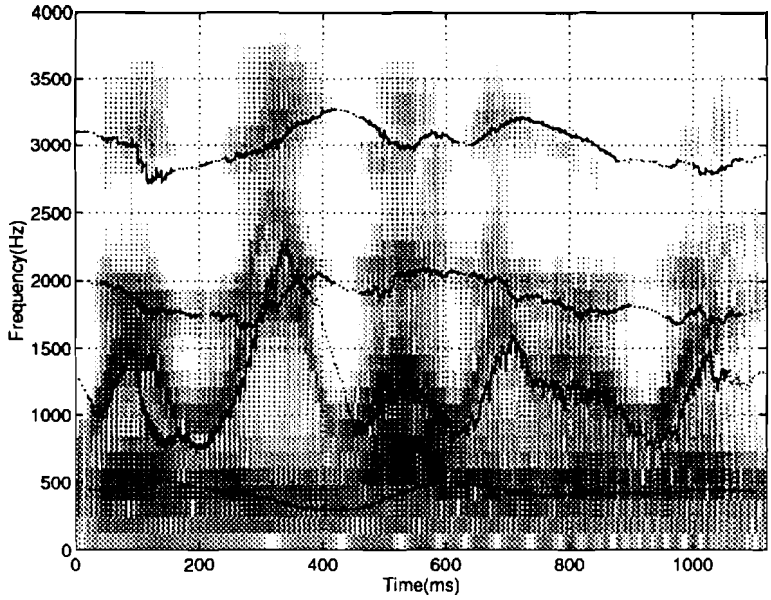


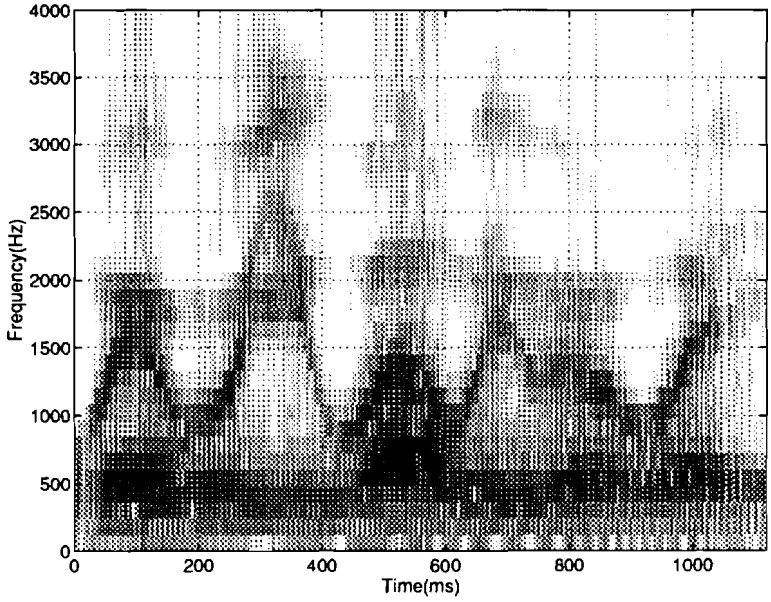
Fig. 3.14. Formant tracks for the stop transition of the word "c/u/ps": $\hat{f}_1(k)$ (lower curve) and $\hat{f}_2(k)$ (upper curve).

3.4 Formant Tracking: An All Voiced Sentence

In this section we apply the model (Eqs. (2.1)–(2.5)) 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 [24, 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, $a_{,,} = a_{,,} = .99$, $q_{\nu_i} = 12$, $p_{f_i,0} = 0$, ancl $p_{\phi_i,0} := 0$. The values of q_{a_i} ancl q_{f_i} vary from formant to formant: $q_{a_i} = 50, 30, 10, 1$; and $q_{f_i} = \sqrt{.5}, \sqrt{22}, 2, 2$ for $i = 1, 2, 3, 4$ respectively. Finally, $r = \sqrt{1/12}$. The spectrogram of the original speech with superimposed plots of the estimates $f_i(k)$ is shown in Figure 3.15(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 3.15(a), the formant tracks extend through regions of



(a)



(b)

Fig. 3.15. The sentence "Where were you while we were away." (a) Original spectrogram and estimated formant tracks. (b) Reconstructed spectrogram.

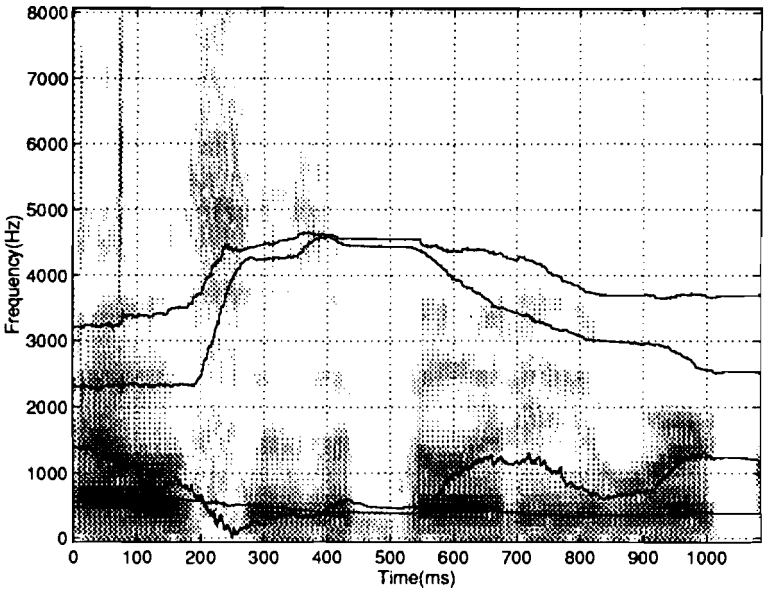
the spectrogram where there is little energy because at sample k we plot the i th formant track $\hat{f}_i(k)$ even when the energy in the i th formant (essentially the energy in $a_i(k)$) is small. To show the distinction, the signal has been divided into frames of duration 20 ms and the i th formant track for a particular frame is plotted as a solid line if and only if the square root of the average energy in $\hat{a}_i(k)$ in the frame is greater than .3 times the standard deviation of $\hat{a}_i(k)$ computed over the entire sentence. Figure 3.15(a) demonstrates good 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 3.15(b) we show the spectrogram of $\hat{y}(k)$ which is very similar to the: spectrogram of $y(k)$ shown in Figure 3.15(a).

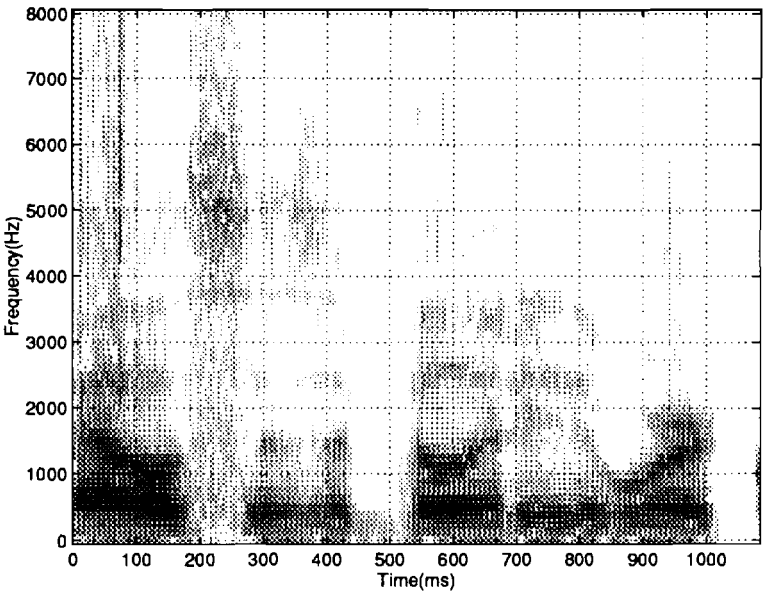
3.5 Application To Mixed Voiced-Unvoiced Speech

In this section we apply the model (Eqs. (2.1)–(2.5)) and EKF to a portion of speech that contains both voiced and unvoiced phonemes. The speech is "Alice's ability to work" from the TIMIT database [24, dr1/msjs1/sx279]. (The sentence is cut short to approximately 1 s). Since it is spoken by the same speaker as the all-voiced sentence in Section 3.4, we use the same parameters except for initial conditions $m_{f_i,0}$ which are 560, 1400, 2300, and 3200 Hz for i = 1, 2, 3, and 4 respectively. The spectrogram of the original speech with superimposed plots of the estimates $\hat{f}_i(k)$ is shown in Figure 3.16(a). The formant tracks lose their interpretation as format frequencies at the unvoiced phoneme /s/ arouhd 200 ms. Both $\hat{f}_3(k)$ and $\hat{f}_4(k)$ increase rapidly driven by the energy concentration around 5000 Hz while $\hat{f}_1(k)$ and $\hat{f}_2(k)$ extend through regions that seem to have the second largest energy concentration in the phoneme. After the plosive /b/, $\hat{f}_2(k)$ tracks the second formant.

From the EKF outputs we compute the reconstructed speech signal $\hat{y}(k)$. In Figure 3.16(b) we show the spectrogram of $\hat{y}(k)$ which is very similar to the spectrogram of $y(k)$ shown in Figure 3.16(a). Therefore, even though the formant interpretation

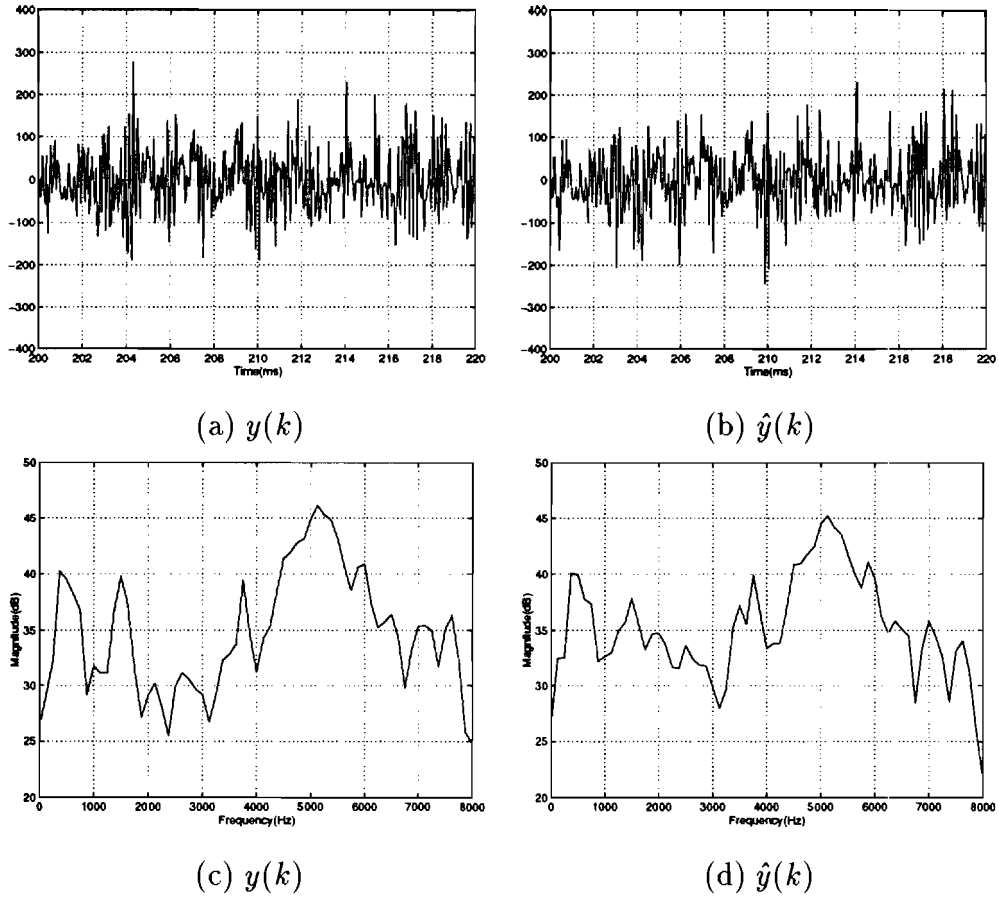


(a)



(b)

Fig. 3.16. The speech "Alice's ability to work". (a) Original spectrogram and estimated formant tracks. (b) Reconstructed spectrogram.

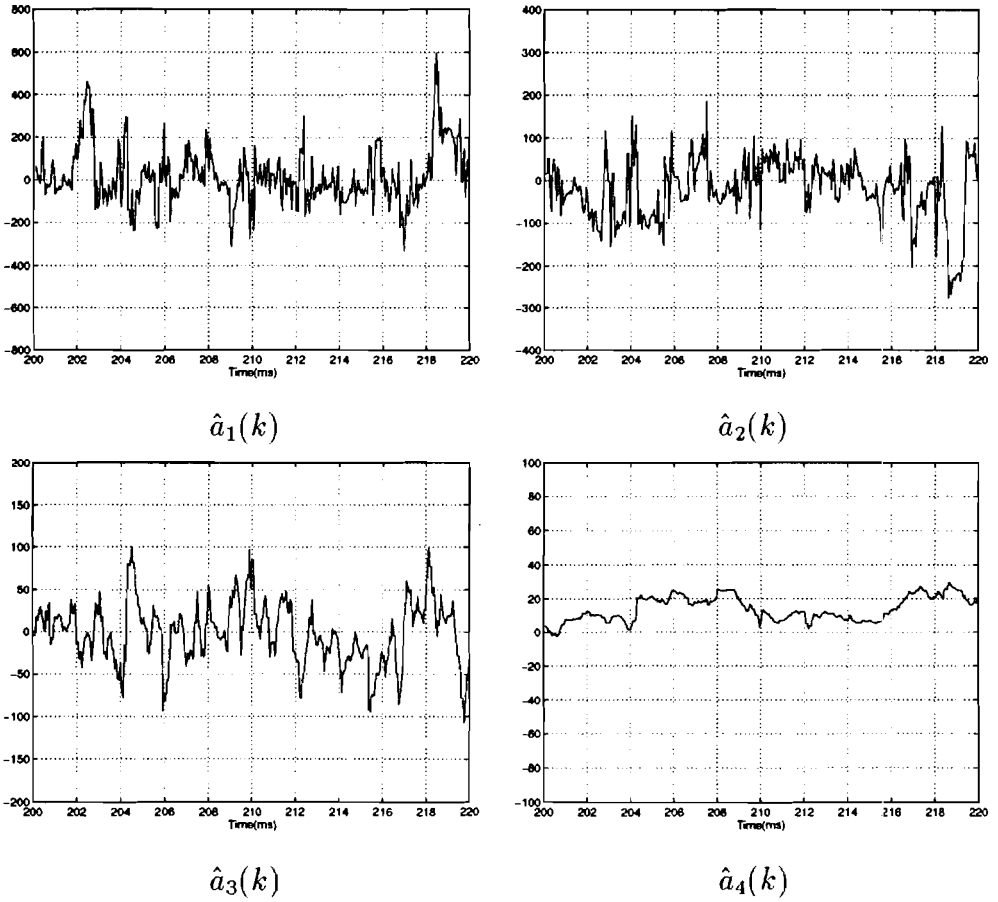


(a) $y(k)$ (b) $\hat{y}(k)$

(c) $y(k)$ (d) $\hat{y}(k)$

Fig. 3.17. The original and reconstructed unvoiced phoneme /s/. (a) and (b): time domain. (c) and (d): frequency domain.

of $a_i(\mathrm{k})\cos(\phi_i(k))$ breaks down in unvoiced phonemes, the the superposition of the $a_i(k)\cos(\phi_i(k))$ signals accurately represents the speech. In Figure 3.1'1 we show the original and reconstructed unvoiced phoneme /s/ in the time and frequency domains while in Figures 3.18 and 3.19 we show the EKF estimates. These estimates seem to be unstructured and mostly random as was also observed in [4, Figure 8].



$\hat{a}_1(k)$ $\hat{a}_2(k)$

$\hat{a}_3(k)$ $\hat{a}_4(k)$

Fig. 3.18. EKF estimates $\hat{a}_i(k)$ for the unvoiced phoneme /s/.

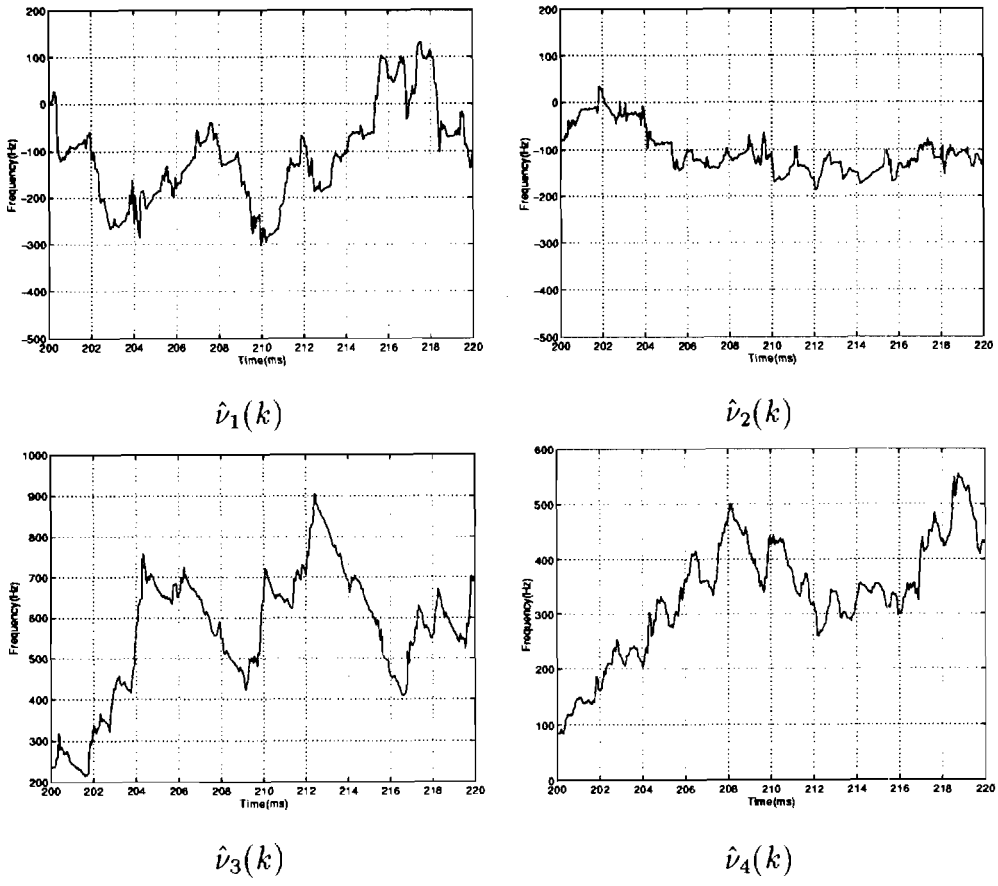


Fig. 3.19. EKF estimates $\hat{\nu}_i(k)$ for the unvoiced phoneme /s/.

4. COMPARISON OF DESA-1 AND MBDA

In this chapter we compare two approaches for extracting the modulating signals from jointly amplitude (AM) and frequency (FM) modulated waveforms: DESA-1, based on Teager's energy operator, and MBDA, based on statistical nonlinear filter. In Section 4.1 we briefly review the two algorithms with a focus on their different characteristics. In the following sections, we apply DESA-1 and MBDA to two examples: a speech phoneme (Section 4.2) and a synthetic two-chirp signal (Section 4.3) [30].

4.1 DESA-1 And MBDA

DESA-1 is a demodulation algorithm [8, 18, 12, 19, 4, 5, 2] that extracts the amplitude and frequency modulations from a jointly AM–FM signal, e.g., a signal modeling a single speech resonance. Let $y(k)$, representing a speech resonance, be modeled as $y(k) = a(k)\cos(\phi(k))$, where $\phi(k) \doteq \Omega_c k + \Omega_m \sum_{n=0}^{k} q(n) + \theta$ for some function $q(\cdot)$ and constants Ω_c, Ω_m, and 0. Define the instantaneous frequency by $\mathrm{Sl}^{\mathrm{inst}}(\mathrm{k}) = \Omega_c + \Omega_m q(k)$. Then the outputs of DESA-1, i.e., the estimates of $|a(k)|$ and $\Omega^{\mathrm{inst}}(k)$, are computed by Eqs. (1.7)–(1.9) which are based on Teager energy operator.

When the Teager energy operator is applied to signals with a superposition of terms, i.e., $y(k) = \sum_i a_i(k)\cos(\phi_i(k))$, or additive noise, i.e., $y(k) = a(k)\cos(\phi(k)) + v(k)$, the operator is applied to each output of a bank of bandpass filters. In the case of a superposition of terms, the bandwidth of the i th filter is determined by the bandwidth of the term $a_i(k)\cos(\phi_i(k))$ and the outputs of the i th energy

operator are estimates of $|a_i(k)|$ and $\Omega_i^{\text{inst}}(k)$. 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 MBDA, a linear superposition of jointly AM–FM terms and the presence of noise are considered simultaneously. MBDA depends on a statistical model for each signal and a simple choice of model [31, 32] is described in Section 2.2. For each formant (i labels the formant), the outputs of MBDA are the estimates of the Kaiser-Teager amplitude signal ($\hat{a}_i$(k)), the Kaiser-Teager frequency signal ($\hat{\nu}_i$(k)), the formant frequency ($\hat{f}_i(k)$), and the total phase signal ($\hat{\phi}_i(k)$) which are extracted from the measured speech signal by extended Kalman filter (EKF). From these estimates we can compute a reconstructed speech signal,denoted by $\hat{y}(k)$, by $\hat{y}(k) = \sum_i \hat{a}_i(k)\cos(\hat{\phi}_i(k))$.

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 $a_i(k)\cos(\phi_i(k))$ 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 (i.e., the bandwidth and center frequencies of the Gabor filters) are seen to qualitatively correspond to the parameters in the statistical model of the nonlinear filtering approach.

Comparison of DESA-1 and MBDA is not easy. $a(k)\cos(\phi(k)) = \text{a}^1(\text{k})\cos(\phi'(k))$ does not imply that $a(k) = \text{a}^1(\text{k})$ and $\phi(k) = \phi'(k)$. It follows that estimation of $a(k)$ and $\phi(k)$ requires prior information in addition to the measurements. Since the prior inforination in DESA-1 and MBDA is quite different, it is not surprising that the results can be quite different. Rather than comparing the estimates of $a(k)$ and $\phi(k)$ from different methods, one might instead compare the speech signals that result from reconstruction using the different estimates. While this is straightforward for MBDA, it is not straightforward for DESA-1 because in the latter approach the

estimate is of $|a(k)|$ rather than $a(k)$ so a sign is lost.

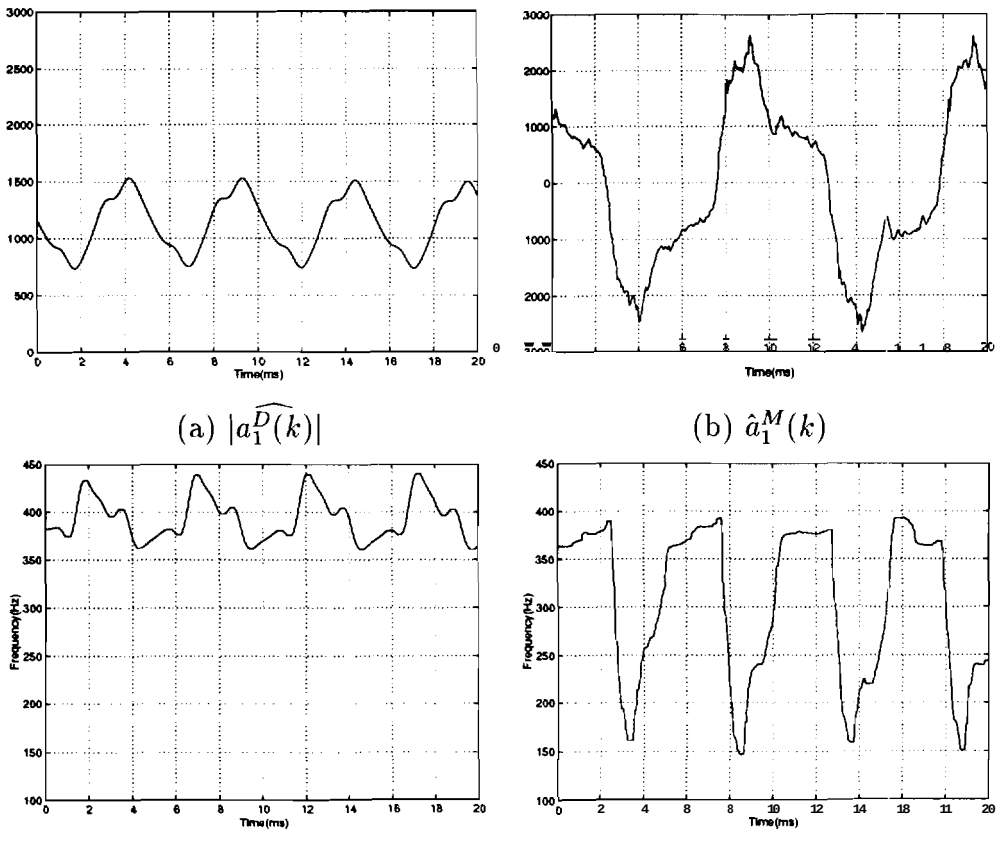
4.2 The Phoneme /ee/

In this section we compare DESA-1 and MBDA on the phoneme /ee/ of the word "m/ee/ting" processed in Section 3.2. For MBDA, we use the same parameters as in Section 3.2. For DESA-1 we extract the first two formants using Gabor filters with center frequencies of 390 Hz and 2200 Hz, $\alpha = 1000$, and N = 55[4]. The original signal in the time and frequency domains and the reconstructed signal from MBDA in the time and frequency domains are shown in Figures 3.9 and 3.10. The MBDA reconstructions are excellent. Because DESA-1 estimates $|a(k)|$ rather than $a(k)$, it is not possible to compute a reconstruction for DESA-1. We also show the instantaneous amplitude estimates $|\widehat{a_i^D(k)}|$ and $\hat{a}_i^M(k)$ for $i = 1, 2$, where superscripts D and M refer to DESA-1 and MBDA respectively. The MBDA estimates are the same as in Figure 3.11 and shown here for comparison. The DESA-1 estimates tend to be smoother than the MBDA estimates. In both cases, structure within the pitch period is visible. In the MBDA case, alternative pulses tend to have reversed signs. Finally, we show the instantaneous frequency estimates $\text{Med}_7(\widehat{\Omega_i^{\text{inst}}(k)})$ and $\text{Med}_7(\hat{f}_i(k) + \hat{\nu}_i(k))$ for $i = 1, 2$ from DESA-1 and MBDA respectively where Med_7 indicates a 7 point median filter [4, Figure 6]. For the first bandpass filter (DESA-1) or formant (MBDA) there is substantial structure within the pitch period. For higher order formants (e.g., 2) the MBDA estimates tend to be relatively unstructured.

4.3 A Two-Chirp Signal

In this section we compare DESA-1 and MBDA on a double chirp signal patterned after the single chirp signal of Ref. [4, Figure 2]. Two signals are considered. The first signal is the same $y(k)$ used in Section 3.1 where it was defined by:

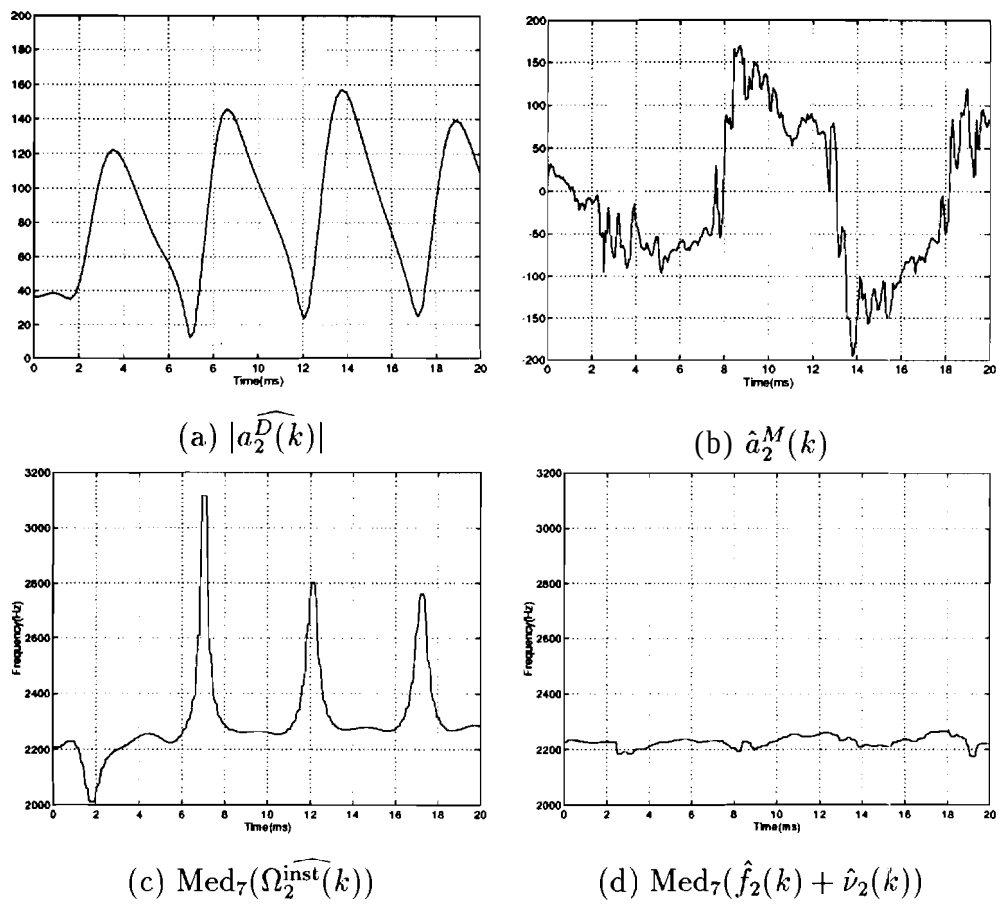
$$\begin{aligned} y(k) &= \cos(2\pi f_a kT)\cos(2\pi(f_{m1} + f_c kT)kT) \\ &\quad + 0.2\cos(2\pi f_a kT)\cos(2\pi(f_{m2} - f_c kT)kT); \end{aligned}$$



(a) $|\widehat{a_1^D(k)}|$ (b) $\hat{a}_1^M(k)$

(c) $\mathrm{Med}_7(\widehat{\Omega_1^{\mathrm{inst}}}(\mathrm{k}))$ (d) $\mathrm{Med}_7(\hat{f}_1(\mathrm{k}) + \hat{\nu}_1(k))$

Fig. 4.1. Phoneme /ee/: first formant.



(a) $|\widehat{a_2^D(k)}|$ (b) $\hat{a}_2^M(k)$

(c) $\mathrm{Med}_7(\widehat{\Omega_2^{\mathrm{inst}}(k)})$ (d) $\mathrm{Med}_7(\hat{f}_2(k) + \hat{\nu}_2(k))$

Fig. 4.2. Phoneme /ee/: second formant.

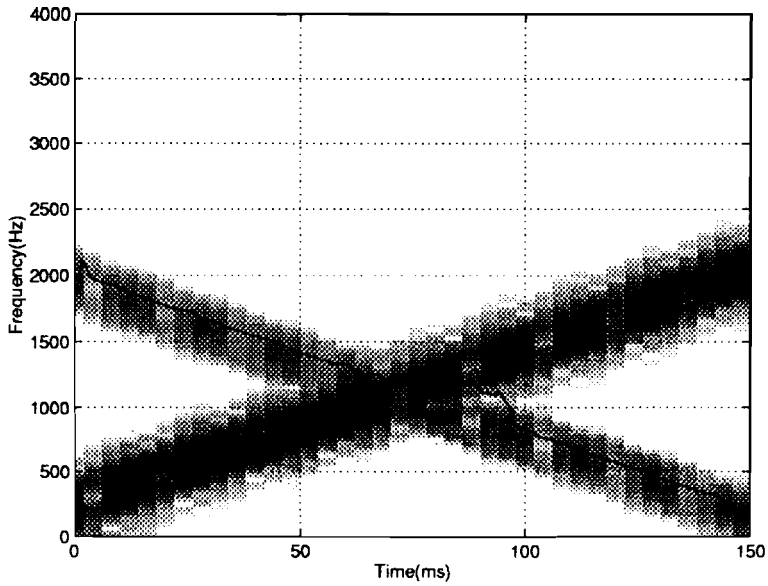
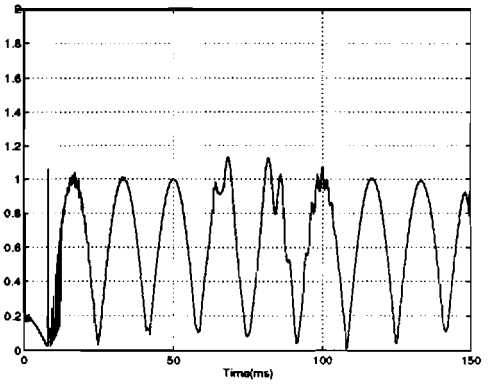


Fig. 4.3. Spectrogram of noise free chirp with $\hat{f}_1(k) + \hat{\nu}_1(k)$ and $\hat{f}_2(k) + \hat{\nu}_2(k)$.

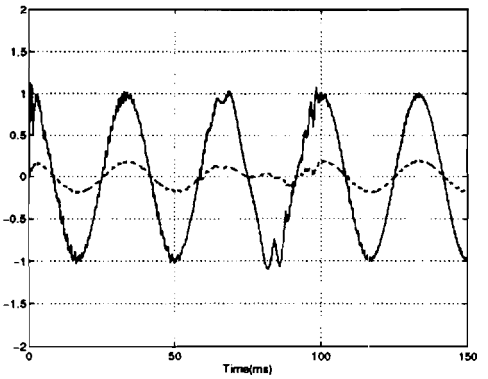
where T = 1/16000 s, f_a = 30 Hz, f_{m1} = 200 Hz, f_{m2} = 2000 Hz, and f_c = 6000 Hz/s. The second signal is $y(k)$ plus additive white Gaussian noise with standard deviation .1778 = $\sqrt{10^{-1.5}}$ = 15 dB. The DESA-1 has 5 bandpass filters with the specifications of Ref. [8, Figure 9]. The MBDA has the same parameters as in Section 3.1 for the noise free case and for the noisy case the observation noise standard deviation r is set to be 1.778 = $\sqrt{10^{-1.5}}$. Ideal performance would lead to $\hat{a}_1(k) = \cos(2\pi f_a kT)$, $\hat{f}_1(k) = f_{m1} + 2f_c kT$, $\hat{a}_2(k) = 0.2\cos(2\pi f_a kT)$, $\hat{f}_2(k) = f_{m2} - 2f_c kT$, and $\hat{\nu}_1(k) = \hat{\nu}_2(k) = 0$.

The spectrogram of the noise free double chirp signal with superimposed plots of the estimates $\hat{f}_1(k) + \hat{\nu}_1(k)$ and $\hat{f}_2(k) + \hat{\nu}_2(k)$ are shown in Figure 4.3. (The spectrogram is computed in the same way as in Section 3.4.) In Figure 4.4 we show $|\widehat{a_1^D(k)}|$, $\hat{a}_1^M(k)$ (solid curve) and $\hat{a}_2^M(k)$ (dashed curve), $\text{Med}_7(\widehat{\Omega_1^{\text{inst}}}(k))$, and $\hat{f}_1(\text{k}) + \hat{\nu}_1(\text{k})$ (solid curve) and $\hat{f}_2(k) + \hat{\nu}_2(k)$ (dashed curve) for the noise free chirp signal. The corresponding plots for the noisy chirp are shown in Figure 4.5.

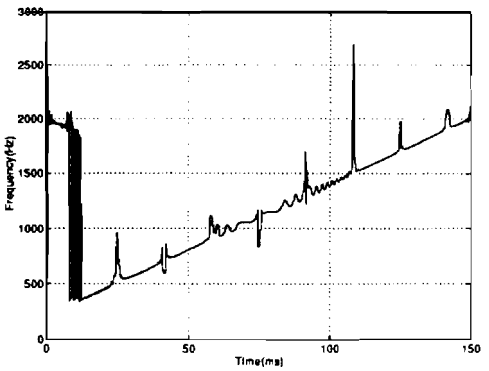
If $x(k)$ is the true signal, $\hat{x}(k)$ is the estimate, and the signal is N samples in du-



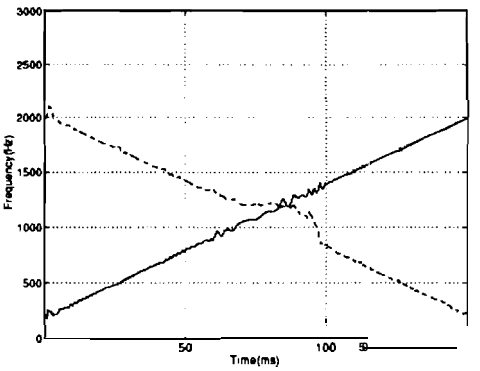
(a) $|\widehat{a_1^D(k)}|$



(b) $\hat{a}_1^M(k)$ (solid curve) and $\hat{a}_2^M(k)$ (dashed curve)

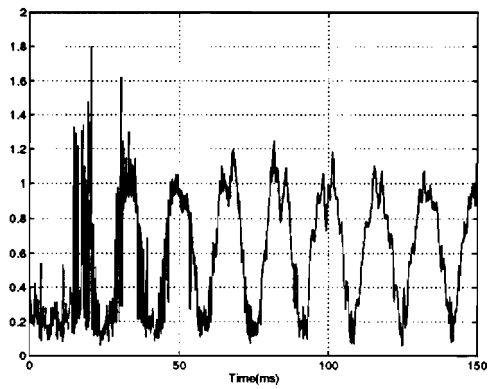


(c) $\mathrm{Med}_7(\widehat{\Omega_1^{\mathrm{inst}}(k)})$

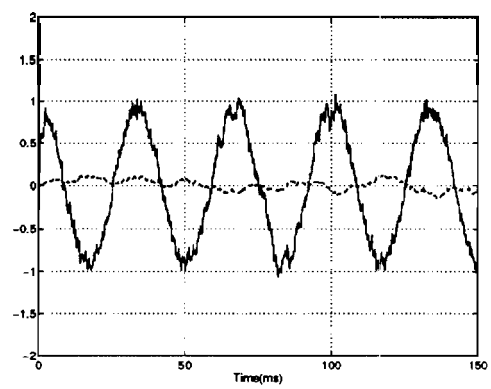


(d) $\hat{f}_1(\mathrm{k}) + \hat{\nu}_1(\mathrm{k})$ (solid curve) and $\hat{f}_2(k) + \hat{\nu}_2(k)$ (dashed curve)

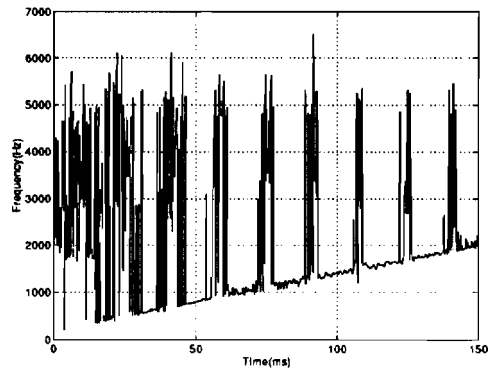
Fig. 4.4. Noise free chirp.



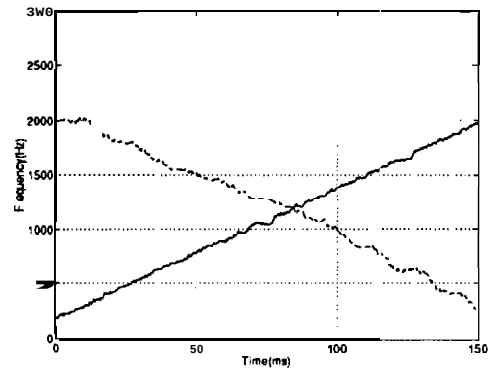
(a) $|\widehat{a_1^D(k)}|$



(b) $\hat{a}_1^M(k)$ (solid curve) and $\hat{a}_2^M(k)$ (dashed curve)



(c) $\mathrm{Med}_7(\widehat{\Omega_1^{\mathrm{inst}}(k)})$



(d) $\hat{f}_1(k) + \hat{\nu}_1(k)$ (solid curve) and $\hat{f}_2(k) + \hat{\nu}_2(k)$ (dashed curve)

Fig. 4.5. Noisy chirp.

Table 4.1.
Mean square error for the two-chirp signal.

	MBDA				DESA-1	
	$\|a_1(k)\|$	$\|a_2(k)\|$	$f_1(k) + \nu_1(k)$	$f_2(k) + \nu_2(k)$	$\|a_1(k)\|$	$\Omega_1^{\text{inst}}(k)$
No noise	0.0034	0.0027	237.57	6734.8	0.0207	$2.0403 \bullet 10^5$
15 d.B noise	0.0154	0.0086	669.14	$1.7696 \bullet 10^4$	0.0470	$3.5770 \bullet 10^6$

ration then in Table 4.1 we report the mean square error performance $\sum_{k=0}^{N-1}(\hat{x}(k) - x(k))^2/N$. In DESA-1 we treat the low energy chirp as noise and only compute one instantaneous amplitude and frequency while in MBDA we compute two. Whether the user regards the low energy chirp as noise or as a second signal is application dependent and in the first instance DESA-1 may be more attractive while in the second instance the energy-tracking ideas of DESA-1 [8] would have to be generalized and MBDA may be more attractive. In terms of mean square error, MBDA performs better than DESA-1. From the plots, most of the error in $\Omega_1^{\text{inst}}(k)$ occurs at tirnes when $a_1(k) = \cos(2\pi f_a kT)$ goes through a zero and DESA-1 therefore selects the band with the greatest noise energy which is usually the highest frequency band because that band is broadest. More sophisticated logic in the energy-tracking algorithm would probably cure this problem.

5. SPEECH CODING

The purpose of this chapter is to document the results of the nonlinear speech coding ideas we proposed in Chapter **1.** We will also discuss some alternative speech coding schemes, such as schemes based on baseband coding and schemes based on subband coding ideas.

5.1 MBDA-Style Coding Idea

The basic MBDA-style coding idea is to use the statistical model and nonlinear filter proposed in previous chapters and compute estimates of $a_i(k)$ and $\phi_i(k)$. Then these estimates are coded, transmitted, and decoded. The speech is reconstructed by combining the decoded estimates through the standard AM-FM nonlinearity (Figure **5.1).** If the coding is perfect, we have already shown that the reconstructed speech signal is very close to the original speech in both the time domain and the frequency domain.

In all the experiments involving real speech signals in this study, we have used statistical models with $\mathbf{I} = 4$ formants. However, we found that among the four formants, the first two lowest formants have the most significant energy. The strength of the estimated amplitude signals decreases from low formants to high formants (Figures **3.11** and **3.12).** Signals reconstructed using estimates from the first two resonances alone are very close to the original speech signals. Our casual listening tests show that such signals are very good both in terms of quality and intelligibility. In Figure **5.2,** we show the wideband spectrogram of the reconstructed all vowel sentence processed in Chapter **3** (Figure 3.15). The only difference is that we

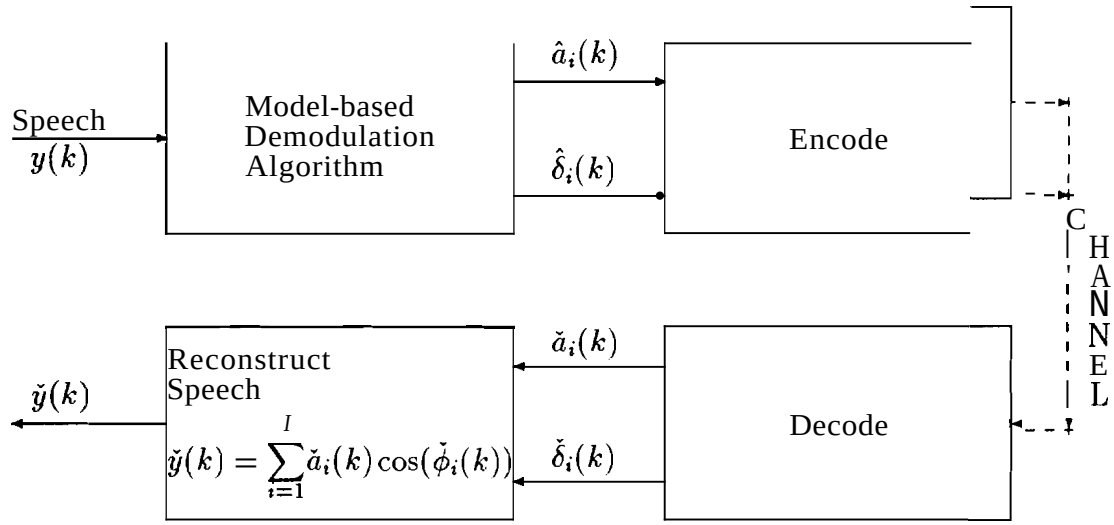


Fig. 5.1. The blockdiagram of MBDA coding

use only $\hat{a}_1$(k), $\hat{a}_2(k)$, $\hat{\phi}_1(k)$, and $\hat{\phi}_2(k)$ when reconstructing $\hat{y}(k)$. From Figure 5.2, we observe that the spectrogram of the reconstructed speech is very similar to the original speech in spite of the fact that only estimates of the first two formants are used. Our casual listening tests support this observation. Thus, to achieve speech coding at a low bit-rate, we will code only first two formants, i.e., we will code $\hat{a}_1(k)$, $\hat{a}_2(k)$, $\hat{\phi}_1(k)$, and $\hat{\phi}_2(k)$.

It follows from the sufficiency of two formants that in MBDA, the term $\hat{a}_i \cos(\hat{\phi}_i)$ not only contains energy over frequencies around f;, but also carries information over a much larger frequency range, especially when i = 1 or 2. This should not be surprising for two reasons: 1) Speech resonances are interwined with each other; and 2) In MBDA, the amplitude and frequency signals are estimated using statistical models simultaneously for all the formants, as opposed to some other demodulation approach (e.g., energy separation algorithm [5, 4]) where the speech signal is first passed through a bank of bandpass filters before processing.

The relationship between $\phi_i(k)$, $f_i(k)$, $\nu_i(k)$, $\hat{\phi}_i(k)$, $\hat{f}_i(k)$, and $\hat{\nu}_i(k)$ are important for coding algorithms. From Eq. (2.4), it follows that

$$\phi_i(k) = \phi_i(0) + 2\pi T \sum_{l=0}^{k-1} (f_i(l) + \nu_i(l)). \tag{5.1}$$

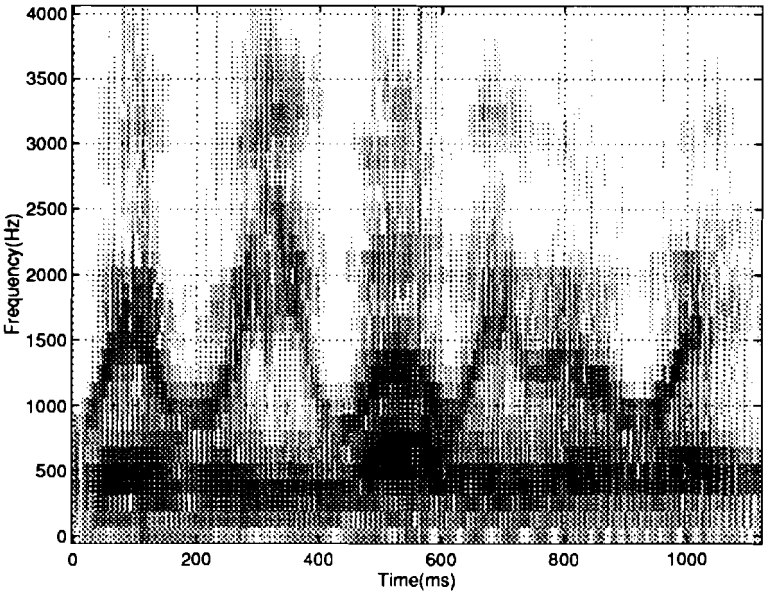


Fig. 5.2. Spectrogram of reconstructed sentence using first two resonances.

There are two natural estimates for $\phi_i(k)$. The first is $\hat{\phi}_i(k)$, which is the estimate of $\phi_i(k)$ from the EKF. The second, denoted by $\bar{\phi}_i(k)$, is defined by

$$\bar{\phi}_i(k) = \bar{\phi}_i(0) + 2\pi T \sum_{l=0}^{k-1} (\hat{f}_i(l) + \hat{\nu}_i(l)), \tag{5.2}$$

where $\hat{f}_i(l)$ and $\hat{\nu}_i(l)$ are the EKF estimates of the formant frequency and the Kaiser-Teager frequency, respectively.

Then,

$$\hat{\phi}_i(k) \approx \mathbf{E}\{\phi_i(k)|y(0),\cdots,y(k-1)\} \tag{5.3}$$

$$= \mathbf{E}\{\phi_i(0) + 2\pi T \sum_{l=0}^{k-1} (f_i(l) + \nu_i(l))|y(0),\ldots,y(k-1)\}. \tag{5.4}$$

On the other hand, from Eq. (5.2), it follows that

$$\bar{\phi}_i(k) \approx \bar{\phi}_i(0) + \mathbf{E}\{2\pi T \sum_{l=0}^{k-1} (f_i(l) + \nu_i(l)|y(0),\cdots,y(l-1)\} \tag{5.5}$$

since $\hat{f}_i(l) \approx \mathbf{E}\{f_i(l)|y(0),\ldots,y(l-1)\}$ and likewise for $\hat{\nu}_i(l)$.

Thus, $\hat{\phi}_i(k) \neq \bar{\phi}_i(k)$ in general because the expectations are conditioned on different observation sequences. Since $\phi_i(k)$ is conditioned on more measurements than $\bar{\phi}_i(k)$, we expect that $\hat{\phi}_i(k)$ is a more accurate estimate of $\phi_i(k)$ than is $\bar{\phi}_i(k)$. This expectation is substantiated in the speech application in the sense that speech reconstructed from $\hat{a}_i(k)$ and $\hat{\phi}_i(k)$ sounds much better than speech reconstructed from $\hat{a}_i(k)$ and $\bar{\phi}_i(k)$. Therefore, we need to code $\hat{\phi}_i(k)$ and not $\hat{\nu}_i(k)$ and $\hat{f}_i(k)$. Because $\phi_i(k)$ is a summation, it is natural to code its first-order difference denoted by $\delta_i(k)$.

5.2 SNR Requirements On Speech Coders

In order to determine the effect of coding $\hat{a}_i(k)$ and $\hat{\delta}_i(k)$ on the reconstructed speech $\hat{y}(k)$ independent of coding methods involved, i.i.d. white Gaussian noise was added to $\hat{a}_i(k)$ and $\hat{\delta}_i(k)$ simultaneously. Let $\acute{a}_i(k)$ and $\acute{\delta}_i(k)$ denote the resulted signals which are defined by

$$\acute{a}_i(k) = \hat{a}_i(k) + \sigma w_{a_i}(\mathrm{k}) \tag{5.6}$$

$$\acute{\delta}_i(k) = \hat{\delta}_i(k) + \sigma w_{\delta_i}(k), \tag{5.7}$$

where $w_{a_i}(k)$ and $w_{\delta_i}(k)$ are i.i.d. , zero-mean, unit variance white Gaussian sequences independent of each other and σ^2 is the noise variance. The reconstructed speech signal is therefore given by

$$\acute{y}(k) = \sum_{i=1}^{2} \acute{a}_i(k) \cos(\sum_{l=0}^{k-1} \acute{\delta}(l)). \tag{5.8}$$

Our casual listening tests indicate that in order for $\acute{y}(k)$ to maintain reasonably good quality, SNR of $\acute{a}_i(k)$ and $\acute{\delta}_i(k)$ must be close to **15** dB.

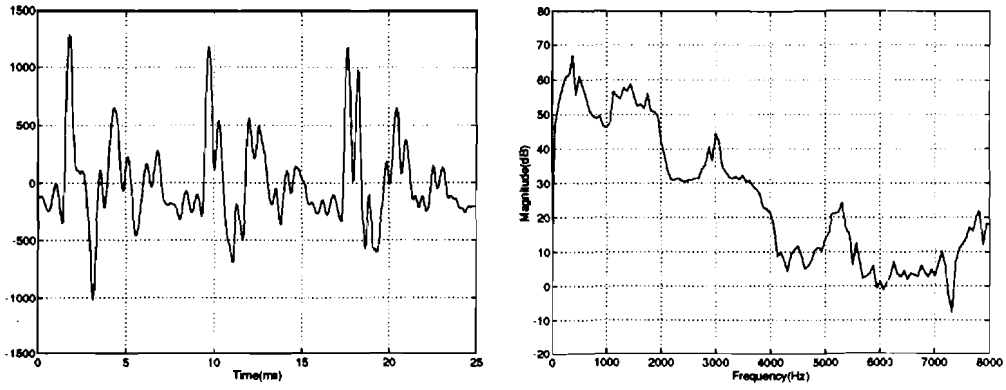
It is obvious that $\hat{a}_i(k)$ and $\hat{\delta}_i(k)$ do not have to be equally well coded since 6; affects the speech quality through the cosine function which is highly nonlinear. Because the speech is a function of $\cos(\phi_i(k))$, it is difficult to control the effect on the speech of errors in coding $\hat{\delta}_i(k)$. Ideally, the noise variance in $\acute{a}_i(k)$ and $\acute{\delta}_i(k)$ should be allowed to be different. However?is is still helpful to use the **15** dB SNR. value as a general guidance in designing coding schemes for $\hat{a}_i(k)$ and $\delta_i(k)$.

5.3 Linear Prediction-Based Coders

We first examine .coders where $\hat{a}_i(k)$ and $\hat{\delta}_i(k)$ are coded using linear prediction ideas. Such ideas have been very successful in coding speech signals, e.g., LPC-10 [33] and CELP [34]. The basic idea is to partition the speech into frames, compute a linear prediction model for the speech within each frame, and then transmit the coefficients of the linear predictive model plus a coded version of the residuals, where the residuals are the difference between the predicted and true speech. Therefore there are two important issues: what order of linear prediction model is required (this controls the number of coefficients that must be transmitted) and whether the residuals can be efficiently coded. In the next subsection we investigate necessary linear predictive model order, which we find to be very low, which makes this approach attractive. Then, in the following two subsections, we examine residual coding using the methods from the LPC-10 and the CELP coders. These results are less attractive: the fixed excitation sequence of LPC-10 is not appropriate for our signals and the codebooks in CELP appear to be too small for our signals because our signals differ more from pitch period to pitch period than do speech signals.

5.3.1 Linear prediction model order

We start with analysis by synthesis techniques using open-loop analysis. For coding $a_i(k)$, we are considering linear predictive coding (LPC) techniques. The remaining examples in this subsection are based on the phoneme /ere/ of the word w/ere/ which we show in the time and frequency domains in Figure 5.3. (All power spectral densities in this section were computed by the Welch method using a 256 point FFT and 50% overlap). In Figure 5.4 we show the residuals of the LPC algorithms with order 1 and 10 applied to the speech signal of Figure 5.3. Notice how non-white the residuals are when the order is 1 in comparison with order 10 (the choice of 10 was motivated by the LPC-10 algorithm [33]). In Figure 5.5 we



(a) (b)

Fig. 5.3. The Phoneme /ere/ of the Word w/ere/. (a) Time domain waveform. (b) F'ower spectral density (Welch method with a 256-point FFT and 50% overlap).

show $\hat{a}_1(k)$ from the EKF and the residuals of LPC predictors of order 1 and 2 applied to $\hat{a}_1(k)$. In Figure 5.6 we show the same quantities for $\hat{a}_2(k)$. Notice how flat these spectra are. Therefore, we believe that LPC encoding at order 1 or 2 will be sufficient. This is a large savings over the standard LPC-10 algorithm because the LPC-10 algorithm uses 41 out of 54 bits/frame for the 10 LPC coefficients [33]. From the time domain waveforms notice that the energy decreases from formant 1 to formant 2 and notice that the impulsive pitch-synchronous behavior seen in the residuals of the LPC-10 algorithm applied to the same speech signal (Figure 5.3) is stronger in formant 1 than formant 2. For higher formants both the energy and the pitch-synchronous behavior decrease further. Therefore: it may be possible to model the residuals for higher formants as i.i.d. Gaussian sequences and code only the variances. In any case, the pitch period is common to all of the signals being coded.

In Figures 5.7 and 5.8 we show the corresponding results for $\delta_1(k)$ and $\delta_2(k)$. In view of these results, we propose to code $\hat{\delta}_i(k)$ using similar LPC techniques. However, in this instance, an LPC of order 0 may be sufficient.

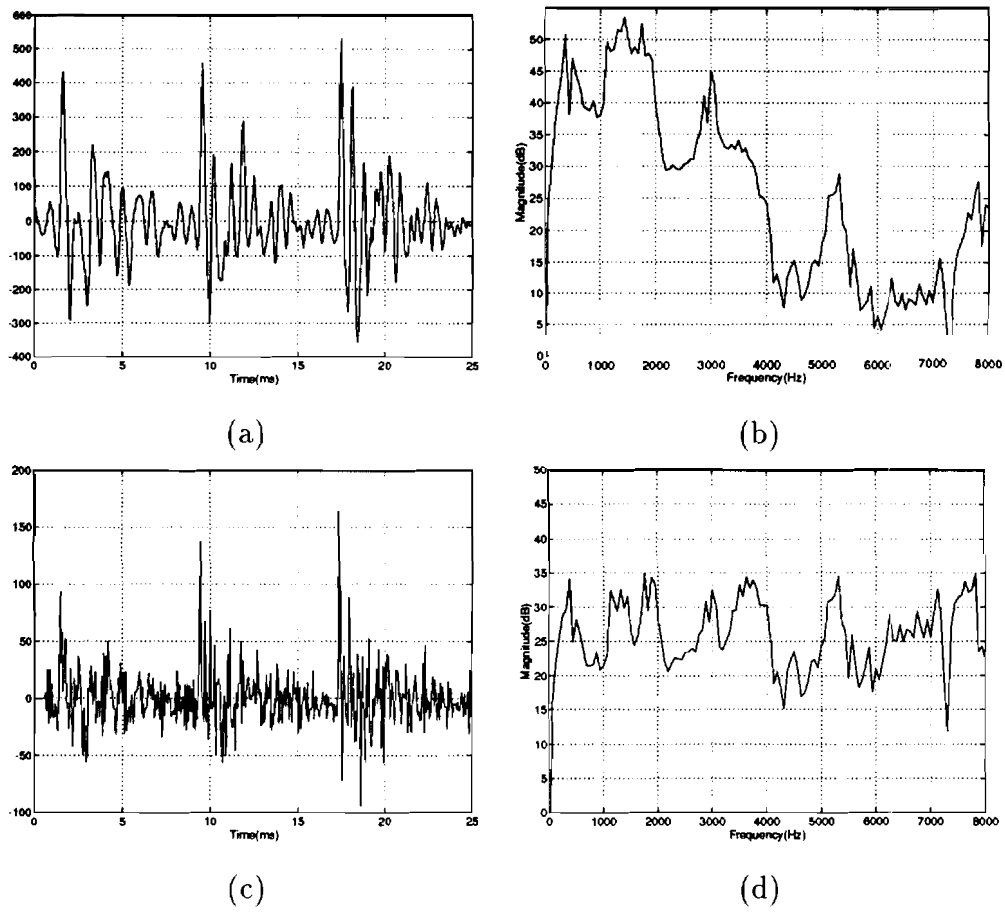


Fig. 5.4. The Residuals from applying an LPC predictor to the speech of Figure 5.3. Residuals from the order 1 LPC predictor: (a) time domain waveform and (b) power spectral density. Residuals from the order 10 LPC predictor: (c) time domain waveform and (d) power spectral density. The power spectral densities were computed by the Welch method with a 256-point FFT and 50% overlap.

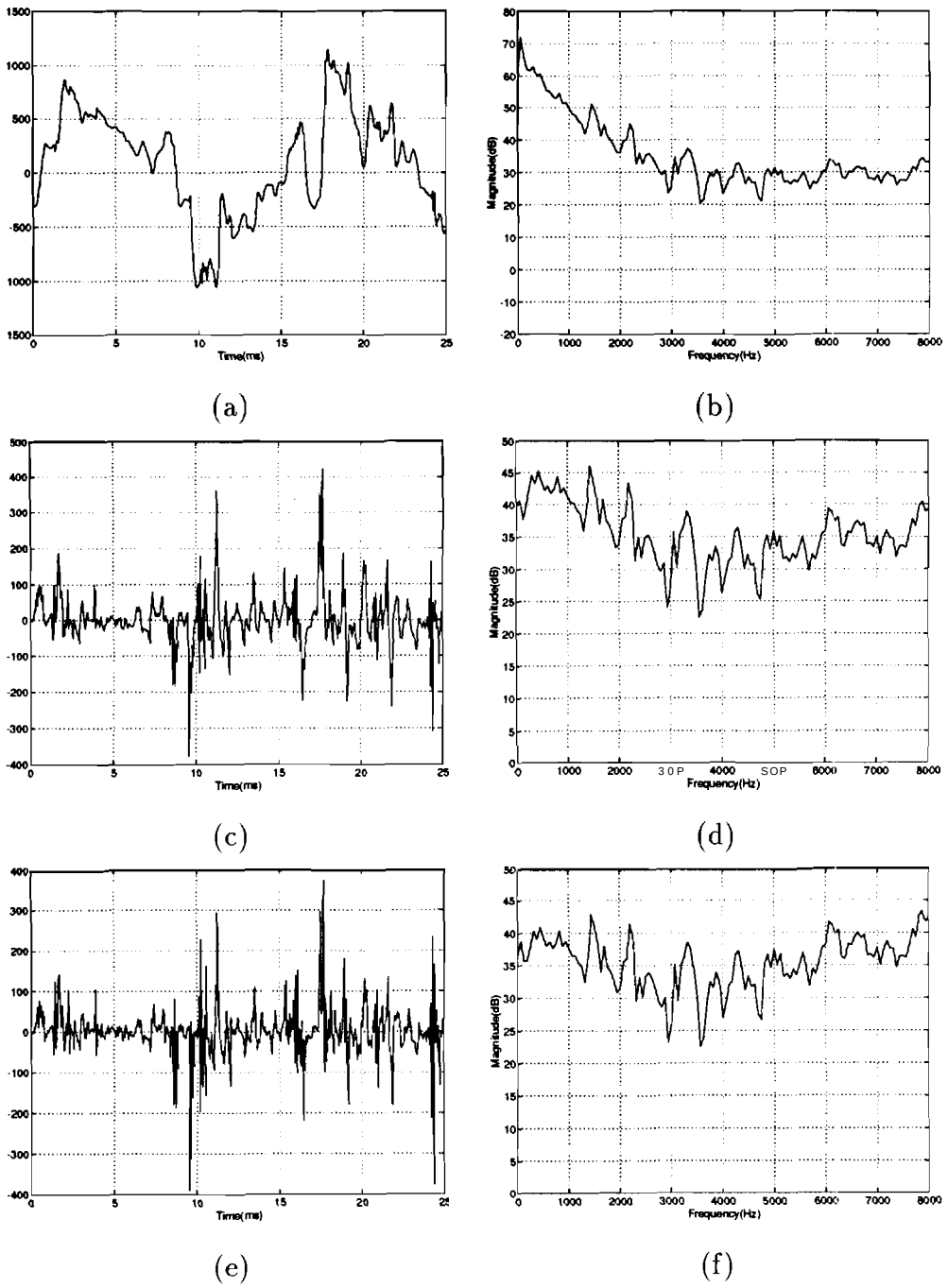


Fig. 5.5. $\hat{a}_1(k)$ and LPC residuals of $\hat{a}_1(k)$ for the speech of Figure 5.3. $\hat{a}_1(k)$: (a) time domain waveform and (b) power spectral density. Residuals from the order 1 LPC predictor of $\hat{a}_1(k)$: *(c)* time domain waveform and (d) power spectral density. Residuals from the order 2 LPC predictor of $\hat{a}_1(k)$: (e) time domain waveform and (f) power spectral density.

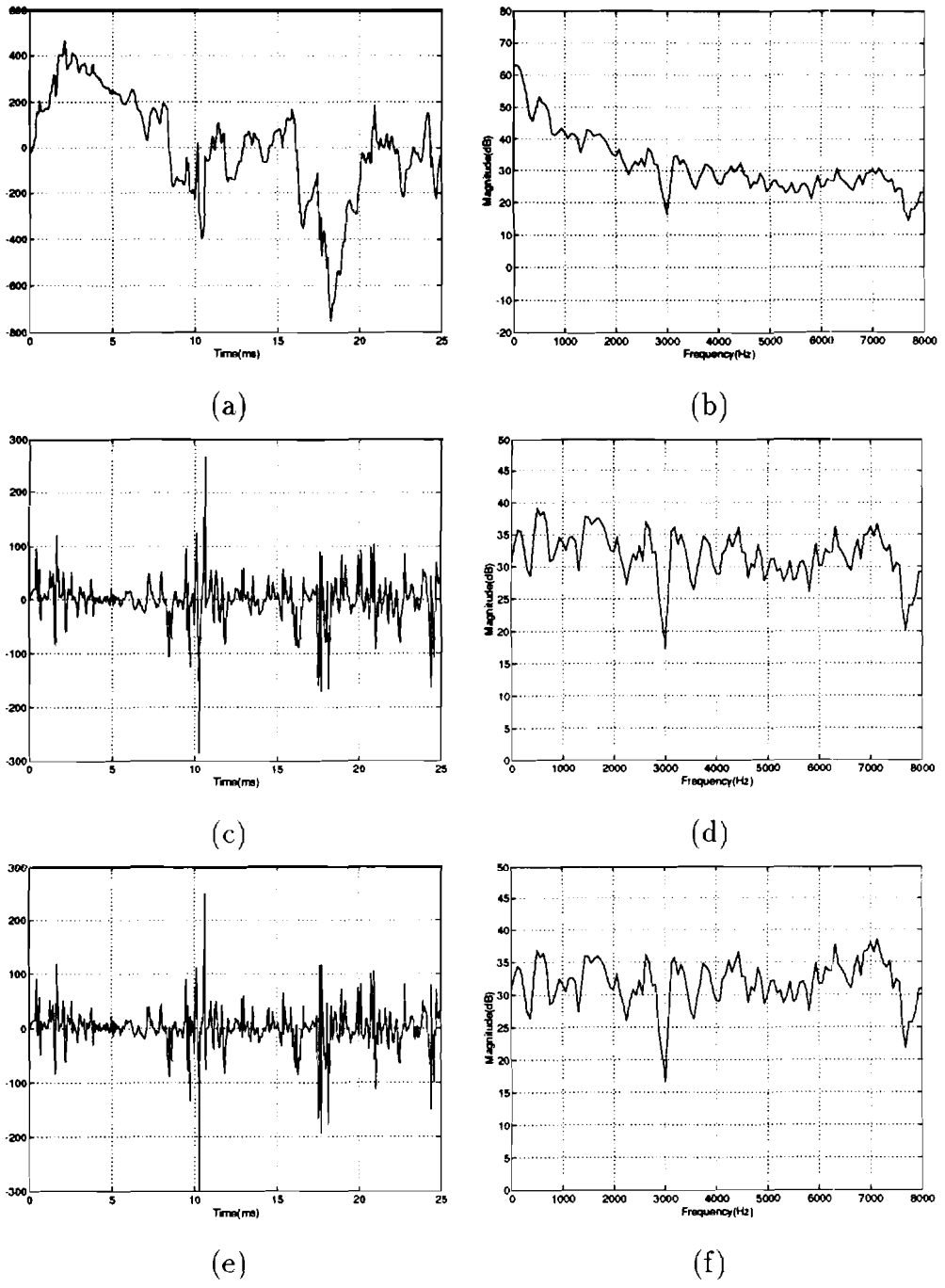


Fig. 5.6. $\hat{a}_2(k)$ and LPC residuals of $\hat{a}_2(k)$ for the speech of Figure 5.3. $\hat{a}_2(k)$: (a) time domain waveform and (h) power spectral density. Residuals from the order 1 LPC predictor of $\hat{a}_2(k)$: (c) time domain waveform and (d) power spectral density. Residuals from the order 2 LPC predictor of $\hat{a}_2(k)$: (e) tirne domain waveform and (f) power spectral density.

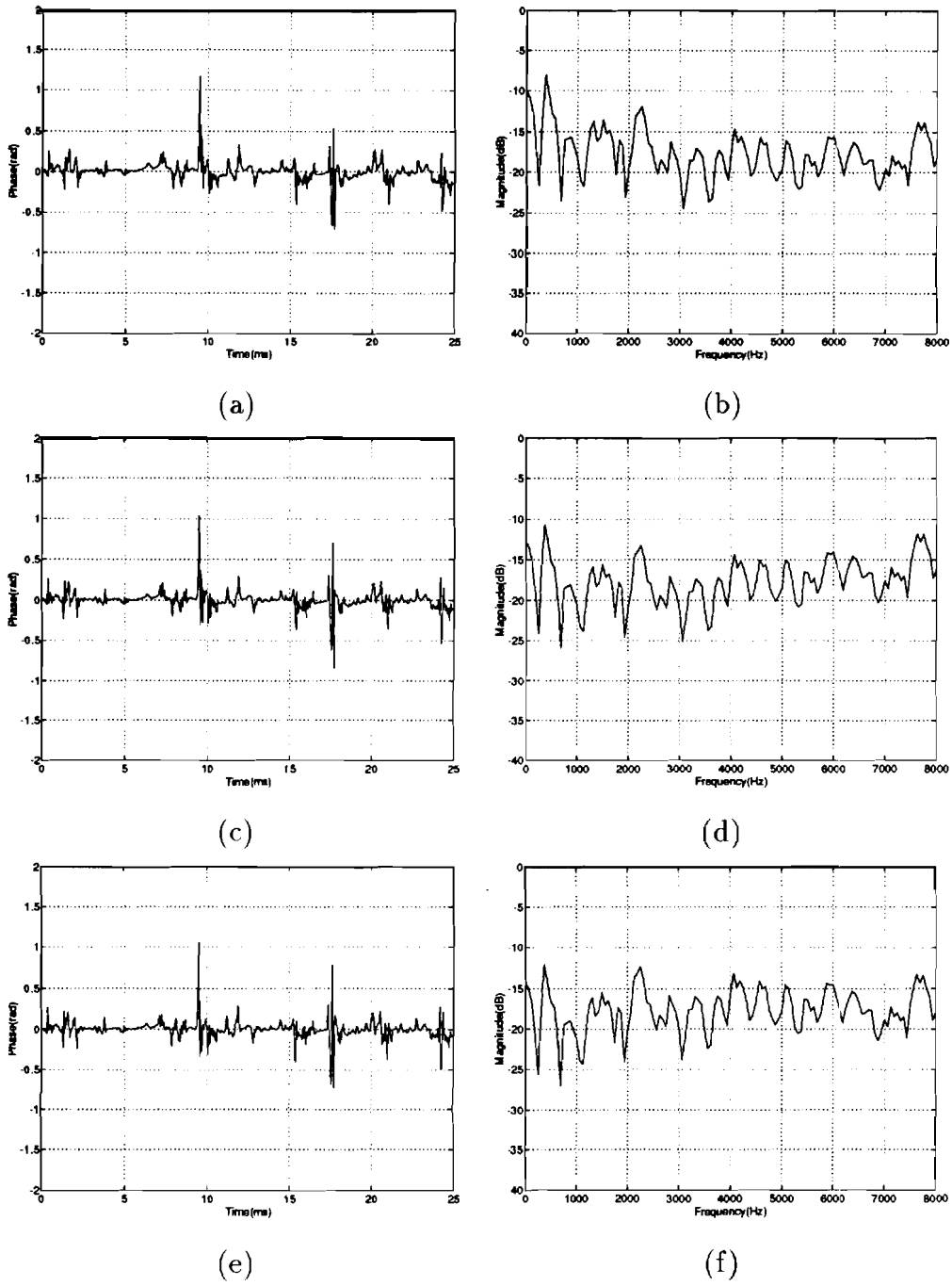


Fig. 5.7. $\hat{\delta}_1(k)$ and LPC residuals of $\hat{\delta}_1(k)$ for the speech of Figure 5.3. $\hat{\delta}_1(k)$: (a) time domain waveform and (b) power spectral density. Residuals from the order 1 LPC predictor of $\hat{\delta}_1(k)$: (c) time domain waveform and (d) power spectral density. Residuals from the order 2 LPC predictor of $\hat{\delta}_1(k)$: (e) time domain waveform and (f) power spectral density.

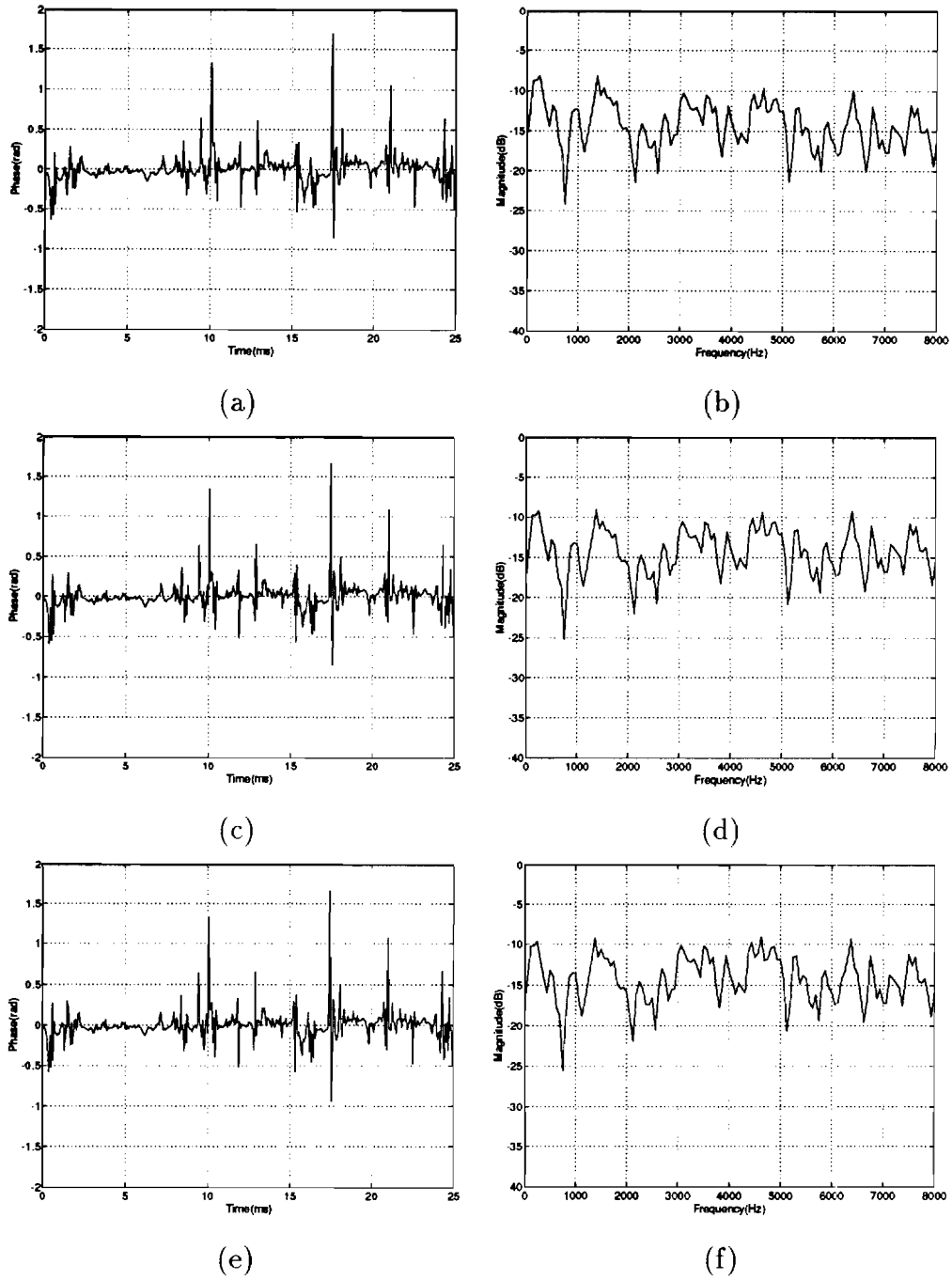


Fig. 5.8. $\hat{\delta}_2(k)$ and LPC residuals of $\hat{\delta}_2(k)$ for the speech of Figure **5.3.** $\hat{\delta}_2(k)$: (a) time domain waveform and (b) power spectral density. Residuals from the order 1 LPC predictor of $\hat{\delta}_2(k)$: (c) time domain waveform and (d) power spectral density. Residuals from the order 2 LPC predictor of $\hat{\delta}_2(k)$: (e) time domain waveform and (f) power spectral density.

5.3.2 MBDA version of the federal standard 1015 (LPC-10)

Federal standard 1015 [33, 35] is a 2.4 kb/s voice coder based on a 10 th linear predictive coding (LPC) model. Therefore, it is also called LPC-10. The input speech is sampled at 8 kHz and is partitioned into 180-sample frames, corresponding to a time duration of 22.5 ms. The 10 th order LPC analysis is performed on each frame. Pitch and voicing are also extracted. 54 bits are transmitted per frame, including 41 bits for the 10 LPC coefficients (for unvoiced speech, only first 4 coefficients are coded, and remaining bits are used for error protection), 7 bits for pitch and voicing, 5 bits for amplitude, and 1 bit for synchronization. Therefore, in LPC-10 more than 70% of the bits are used to transmit the 10 LPC coefficients when the speech is voiced. At the receiver, the pitch/voicing code is used to determine the excitation function to be used. If the speech is unvoiced, pseudorandom numbers are generated and used for the excitation; if the speech is voiced, then a. locally stored waveform, representing one cycle of a plausible prediction residual, is used as the excitation. This stored waveform [36] (Figure 5.9) is 40 samples long; it is truncated or padded out with zeros as required to match the current pitch period.

As we discussed in the previous subsection, in order to code $\hat{a}_i(k)$ and $\hat{\delta}_i(k)$ we propose to adopt standard coding ideas. Specifically, LPC-based techniques are employed to code $\hat{a}_i(k)$ and $\hat{\delta}_i(k)$. In Subsection 5.3.1 we showed that the advantage of this approach is that the order of the LPC model can be dramatically reduced so that fewer LPC coefficients need to be transmitted. When LPC is applied to $\hat{a}_i(k)$, order 2 is sufficient while order 0 is sufficient for $\hat{\delta}_i(k)$. This is a large saving over LPC-10 because the LPC-10 algorithm uses 41 out of 54 bits/frame for the 10 LPC coefficients. It is also worth noticing that the pitch period is common to all four signals, $\hat{a}_i(k)$ and $\delta_i(k)$, $i = 1, 2$.

The bit allocation for our proposed MBDA version of LPC-10 is given below. The sampling rate is 8 kHz and the frame length is 180 samples (22.5 ms).

- $\hat{a}_1(k)$ and $\hat{a}_2(k)$ are coded using order 2 LPC with each coefficient using 5 bits

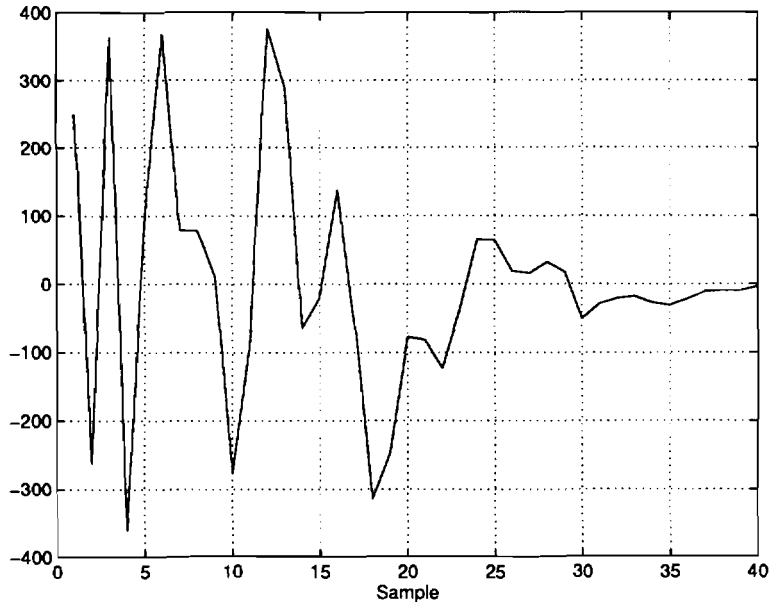


Fig. 5.9. The excitation of voiced speech in LPG-10.

while $\hat{\delta}_1(k)$ and $\hat{\delta}_2(k)$ are coded with order 0 LPC.

Signal	$\hat{a}_1$	$\hat{a}_2$	$\hat{\delta}_1$	$\hat{\delta}_2$
p	2	2	0	0
bits/frame	10	10	0	0

- The same pitch and voicing are used for all 4 signals for a total cost of 7 bits/frame.

- The amplitudes of $\hat{a}_1(k)$ and $\hat{\delta}_1(k)$ are coded using 5 bits/frame, respectively. The amplitude range of $\hat{a}_2(k)$ and $\hat{\delta}_2(k)$ is much decreased relative to $\hat{a}_1(k)$ and $\hat{\delta}_1(k)$ and fewer bits are necessary. The amplitude of $\hat{a}_2(k)$ is coded as 2^{-n} times the amplitude of $\hat{a}_1(k)$ using 2 bits/frame and likewise for $\hat{\delta}_2(k)$.

Signal	$\hat{a}_1$	$\hat{a}_2$	$\hat{\delta}_1$	$\hat{\delta}_2$
bits/frame	5	2	5	2

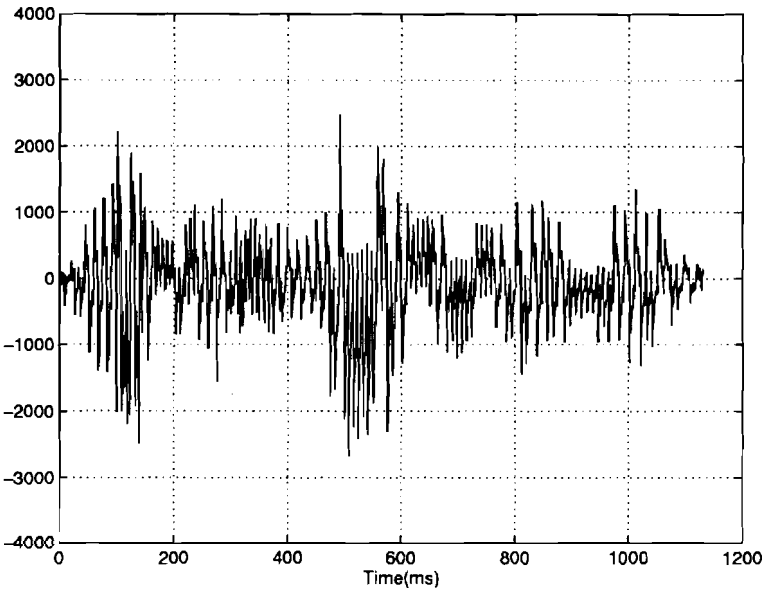
- 1 bit/frame is used for synchronization.

This leads to 42 bits/frame, corresponding to a hit-rate of 1.87 kb/s.

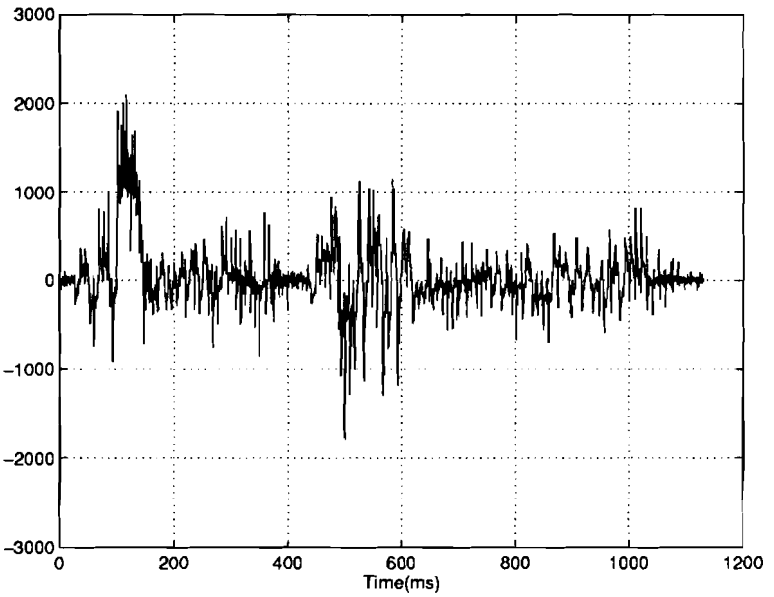
Since there are no MBDA-style coders, we started the implementation of the above ideas by modifying federal standard 1015 software. We first studied the source codes in an effort to understand the function of each program with the help of government publications. Then we modified the codes, paying special attention to the order of the LPC model, perceptual weighting, prefiltering, and postfiltering. The modified software, which is the MBDA version of LPC-10, is then applied to $\hat{a}_1(k)$, $\hat{a}_2(k)$, $\hat{\delta}_1(k)$, and $\hat{\delta}_2(k)$. Figures 5.10 and 5.11 show the four signals for the sentence "Where were you while we were away" processed in previous chapters. The coded–decoded signals are illustrated in Figures 5.12 and 5.13. The SNRs of the signals shown are −2.43 dB, −1.01 dB, −0.13 dB and 0.17 dB for $\hat{a}_1(k)$, $\hat{a}_2(k)$, $\hat{\delta}_1(k)$, and $\hat{\delta}_2(k)$, respectively. A significant amount of distortion has been introduced in the coding of these signals. Overall, the SNRs are far below 15 dB.

The difficulties of the MBDA-style LPC-10 coder could be due to the fact that a prestored waveform (i.e., Figure 5.9) is used as the excitation of the LPC model. While the waveform may be suitable for speech signals, there is, indeed, no justification to use it on non-speech signals such as $\hat{a}_i(k)$ and $\delta_i(k)$.

In an LPC-based voice coder, two aspects are crucial in achieving low bit-rate: small parameterization and pitch synchronous residuals that can be coded with reasonably fidelity. We certainly have achieved the first goal by using an LPC model of dramatically low order. As for the second goal, we do not see how a modified LPC-10 can handle that, mainly because of the nature of the built-in waveform. Rather than attempting to redesign the excitation waveform for each of the different $\hat{a}_i(k)$ and $\hat{\phi}_i(k)$ signals, we instead searched the literature for a standard coder that did not have built-in excitation sequences. Federal standard 1016, a code-excited linear predictive (CELP) voice coder, seems to merit investigation.

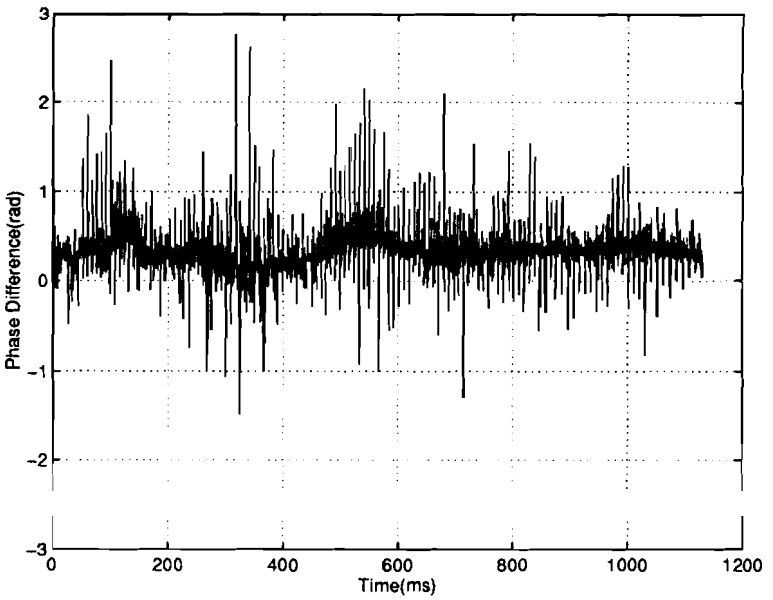


(a) $\hat{a}_1(k)$

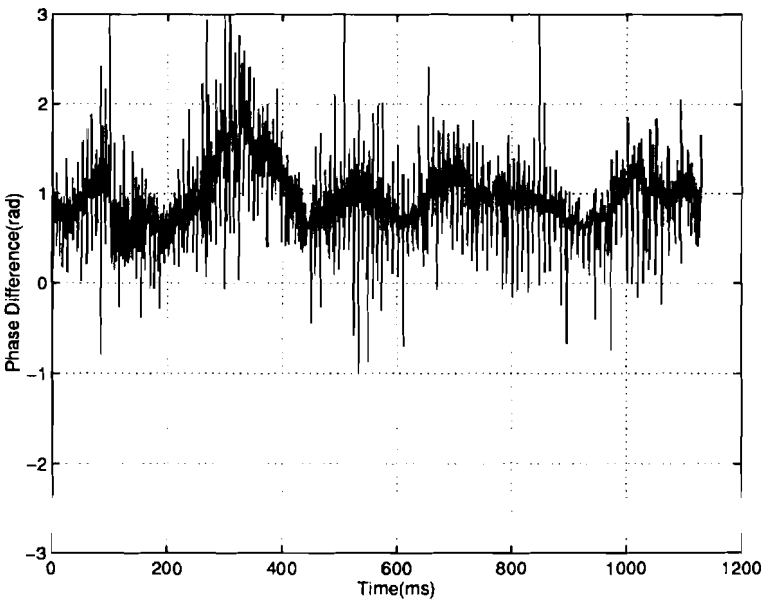


(b) $\hat{a}_2(k)$

Fig. 5.10. The estimated amplitudes for “Where were you while we were away".

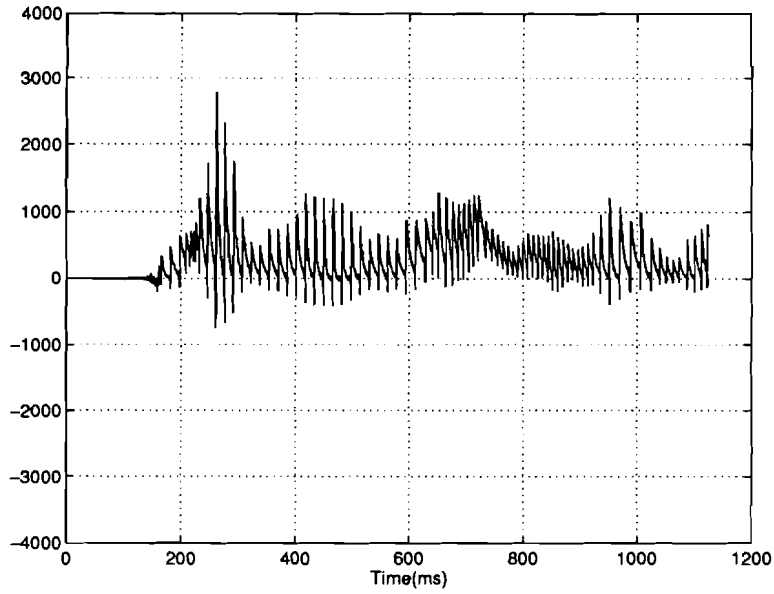


(a) $\hat{\delta}_1(k)$

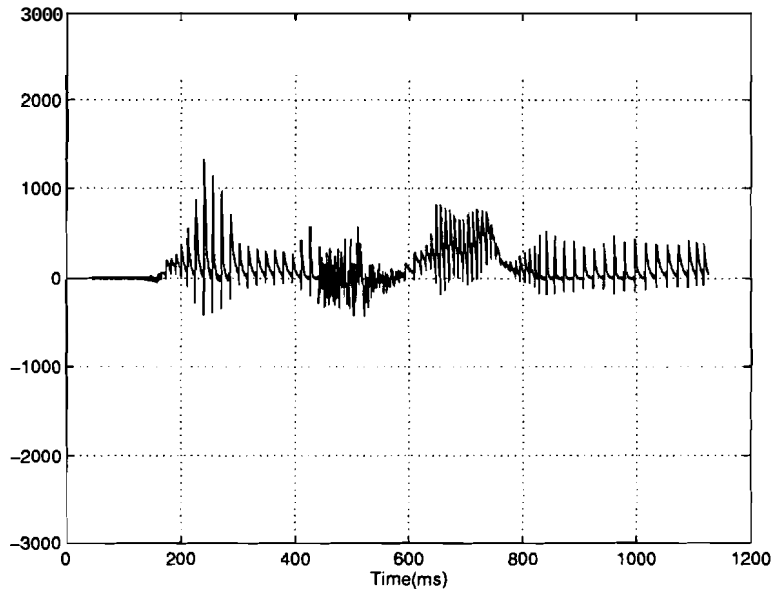


(b) $\hat{\delta}_2(k)$

Fig,.5.11. The estimated phase differences for "Where were you while we were away".

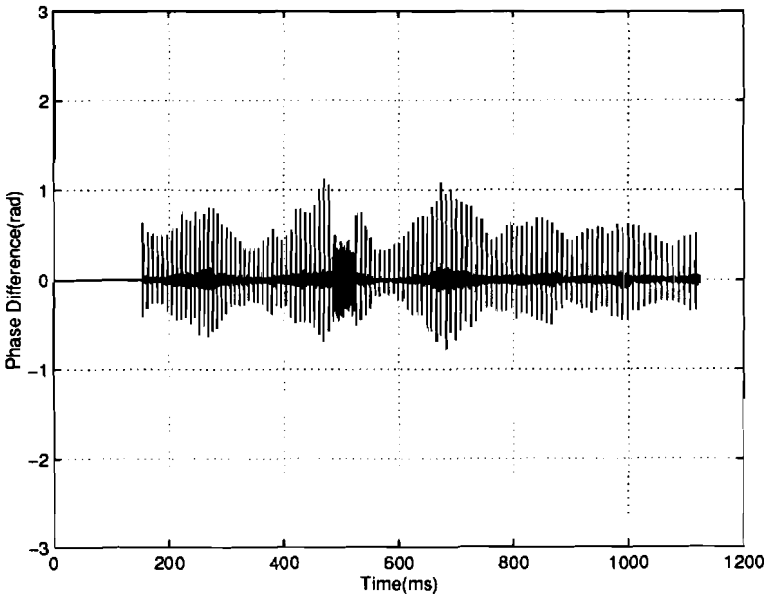


(a) Decoded $\hat{a}_1(k)$

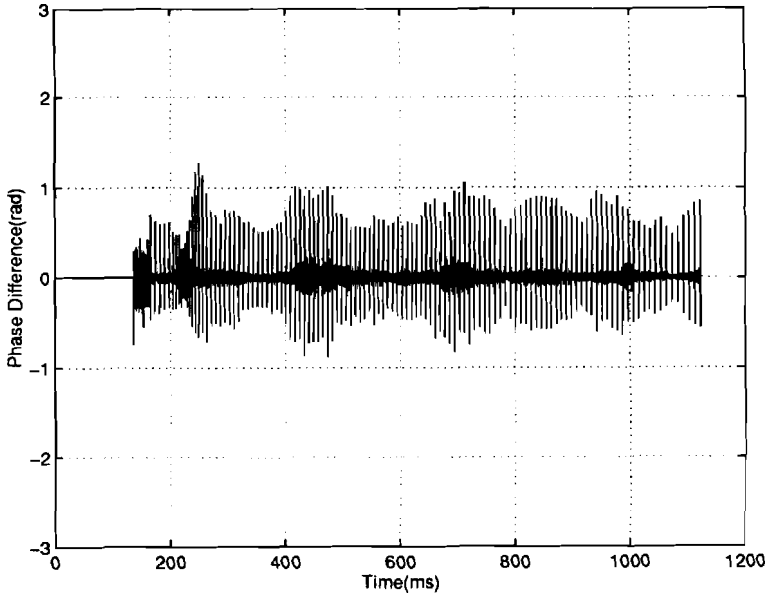


(b) Decoded $\hat{a}_2(k)$

Fig;.5.12. The LPC decoded $\hat{a}_i(k)$ for "Where were you while we were away".



(a) Decoded $\hat{\delta}_1(k)$



(b) Decoded $\hat{\delta}_2(k)$

Fig. 5.13. The LPC decoded $\hat{\delta}_i(k)$ for "Where were you while we were away".

5.3.3 MBDA version of the federal standard 1016 (CELP)

Federal standard 1016 [34, 37, **38,** 39] is a **4.8** kb/s voice coder utilizing code-excited linear prediction (CELP) ideas. Input speech sampled at **8** kHz is segmented into frames of **240** samples (or 30 ms in time) which are processed as one unit. CELP coding is based on analysis-by-synthesis search procedures, perceptually weighted vector quantization (VQ), and linear prediction (LP). **A** 10 th order LIP filter is used to model the speech signal's short-time formant structure. Long-term signal periodicity is modeled by an adaptive codebook VQ (also called pitch VQ). The error from the short-term LP and pitch VQ is vector quantized using a fixed stochastic code'book. The optimally scaled excitation vectors from the adaptive and stochastic code'books are selected by minimizing a time-varying, perceptually weighted distortion measure that improved subject speech quality. The stochastic codebook is ternary valued (-1, 0, +1) and has **512** codewords. The adaptive codebook has **256** codewords.

We modified the CELP software in the same way as we did the LPC-10 software. Specifically, we reduced the order of the LP analysis from **10** to **2** for $\hat{a}_i(k)$ and to 0 for $\hat{\delta}_i(k)$. In addition, some of the perceptually weighting and bandwidth expansion features were removed because the modified CELP is intended to run on EKF estimates rather than on actual speech signals.

The optimal way to use CELP to code MBDA estimates would be to run CELP jointly on $\hat{a}_i$ and δ_i with a closed-loop cost that measures the distortion in $\hat{a}_i(k)\cos(\hat{\phi}_i(k)))$. Howev r, the resulted algorithm is very complicated! and requires a tremendous amount of computation. **A** suboptimal solution is to apply CELP to $\hat{a}_i(k)$ with a closed-loop cost that measures the distortion in $\hat{a}_i(k)$ and independently apply CELP to $\hat{\delta}_i(k)$ with a closed-loop cost that measures the distortion in $\cos(\hat{\phi}_i(k))$ since $\hat{\delta}_i(k)$ contributes to the speech through the cosine function.

Let $x^{(l)}$ be the optimal vector being searched for and g_l be the corresponding optimized gain. Let column vectors $\check{\delta}_i^{(l)}$ and $\check{e}^{(l)}$ denote the decoded $\hat{\delta}_i$ and error signal over a frame as a result of $x^{(l)}$ and g_l. Let $\check{\phi}_i^{(l)}$ be the phase corresponding to

$\check{\delta}_i^{(l)}$. Furthermore, let H denote the LP filter. Then

$$\check{\delta}^{(l)} = Hg_l x^{(l)} = g_l y^{(l)}, \tag{5.9}$$

where $y^{(l)} = Hx^{(l)}$ is the filtered codeword.

If we apply CELP to a cost which measures the distortion in $\cos(\hat{\phi}_i(k))$, then

$$\check{e}^{(l)} = \cos(\hat{\phi}_i) - \cos(\check{\phi}_i^{(l)}) \tag{5.10}$$

$$= \cos(A\hat{\delta}_i) - \cos(A\check{\delta}_i^{(l)}), \tag{5.11}$$

where A is a lower triangular matrix of appropriate size with all I entries, and $\hat{\phi}_i = A\hat{\delta}_i$, $\check{\phi}_i^{(l)} = A\check{\delta}_i^{(l)}$ for column vectors $\hat{\phi}_i$, $\check{\phi}_i^{(l)}$, $\hat{\delta}_i$, and $\check{\delta}_i^{(l)}$ over a frame.

Let $\check{E}^{(l)}$ denote the norm or total square error for codeword 1:

$$\begin{aligned}\check{E}^{(l)} &= \|\check{e}^{(l)}\|^2 \\ &= \cos^T(A\hat{\delta}_i)\cos(A\hat{\delta}_i) - 2\cos^T(A\hat{\delta}_i)\cos(Ag_l y^{(l)}) + \cos^T(Ag_l y^{(l)})\cos(Ag_l y^{(l)}).\end{aligned}$$

Thus,

$$\frac{\partial \check{E}^{(l)}}{\partial g_l} = 2[\cos^T(A\hat{\delta}_i) - \cos^T(Ag_l y^{(l)})]\mathrm{diag}(Ay^{(l)})\sin(Ag_l y^{(l)}). \tag{5.12}$$

No closed-form solution for g_l exists to the equation $\frac{\partial \check{E}^{(l)}}{\partial g_l} = 0$. Numerical solution must be used and the computational requirements of codebook search are very expensive.

Another alternative suboptimal approach is to apply CELP to a cost that measures the distortion in $\hat{\delta}_i(k)$, revise 6; based on the decoded $\hat{\delta}_i$, and apply CELP to the revised $\hat{a}_i$. Let $\tilde{a}_i$ denote the revised $\hat{a}_i$ which is defined by

$$\tilde{a}_i(k) = \hat{a}_i(k)\cos(\hat{\phi}_i(k))/\cos(\check{\phi}_i(k)), \tag{5.13}$$

where $\check{\phi}_i$ is the phase corresponding to the decoded $\hat{\delta}_i$.

Figures 5.14 and 5.15 show the results of modified CELP applied to the signals in Figures 5.10 and 5.11. The corresponding SNRs are 9.66 dB, 10.21 dB, 7.36 dB

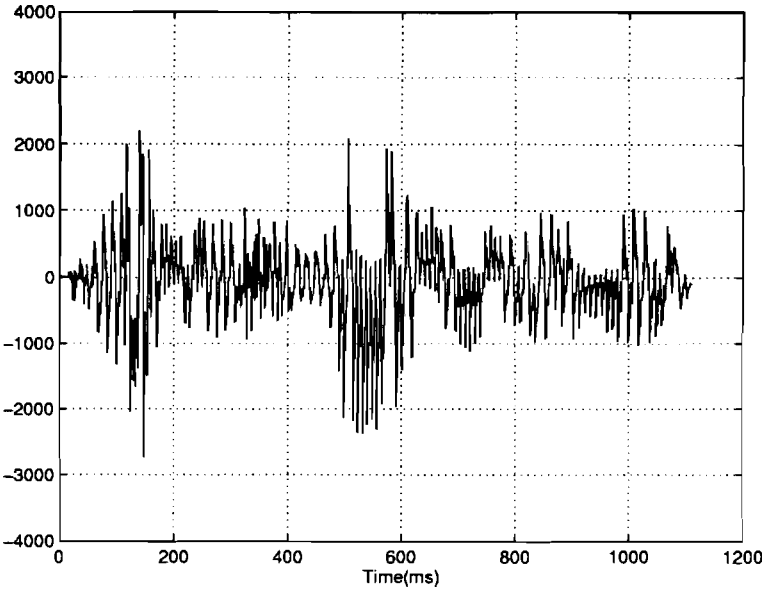
and 13.35 dB for $\hat{a}_1$, $\hat{a}_2$, $\hat{\delta}_1$, and $\hat{\delta}_2$, respectively. While the coding of the two amplitudes has improved substantially, a significant amount of distortion still exists in the coding of the phase signals. This is especially unfavorable due to the fact that $\check{\delta}_i(k)$ needs to be summed up to yield $\check{\phi}_i(k)$ and therefore distortion will be accumulated. Figure 5.16 shows the ratio $\cos(\hat{\phi}_1(k))/\cos(\check{\phi}_1(k))$ for the same sentence. The dynamic range is large. It is evident that some of the coherent structure in $\hat{a}_1$ will be destroyed as a result of the revision and therefore the performance of CELP on the revised signal $\tilde{a}_1$ will decrease. To successfully apply this revision idea, it seems crucial to code the two phase signals with reasonably good fidelity.

We believe that in order for CELP-based techniques to generate reasonably good results on MBDA outputs, relatively large codebooks will be required to represent the LP residuals. This is especially true when the residuals of the two phase signals do not exhibit strong pitch synchronous behavior. But the codebooks of the federal stantlard 1016 and its variations [40, 41] are relatively small and highly structured. The nonlinear cosine function also makes the issue more complicated.

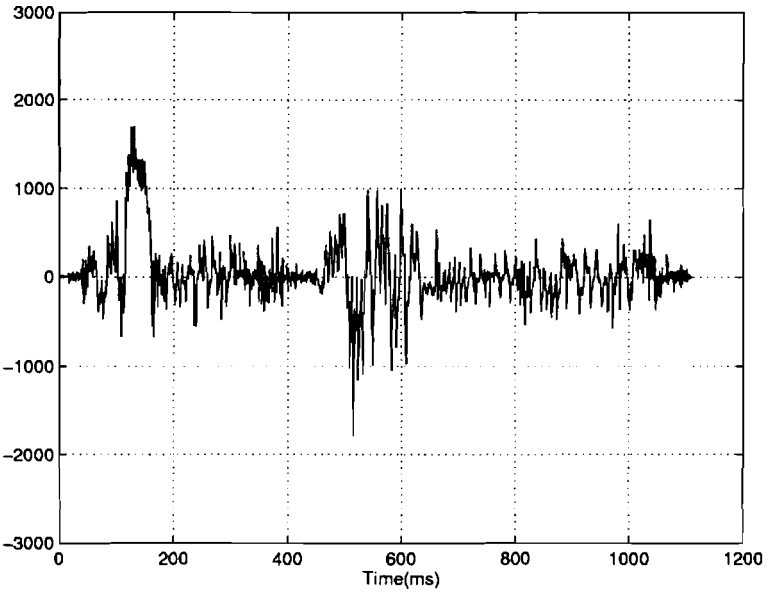
5.4 Other Ideas On Coding MBDA Outputs

We have experimented with several ideas for coding EKF estimates other than the LPC and CELP approaches described above. Our main focus is on the phase signals.

In MBDA, the observation noise has standard deviation r (Eq. (2.5)). Since all the speech data in this study come from TIMIT which is essentially noise free, the observation noise is assumed to be the quantization noise which, under a. uniform $\pm 1/2$-bit model, has standard deviation $\mathrm{r} = \sqrt{1/12}$. With this r, we have been able to process speech in such a way that the reconstructed signal accurately tracks the original speech in both the time domain and the frequency domain. While this is very impressive from the point view of MBDA, it actually poses a problem for coding: the estimated signals vary rapidly in time. If r is increased, the nonlinear estimator would tend to attribute more variations to the observation noise rather

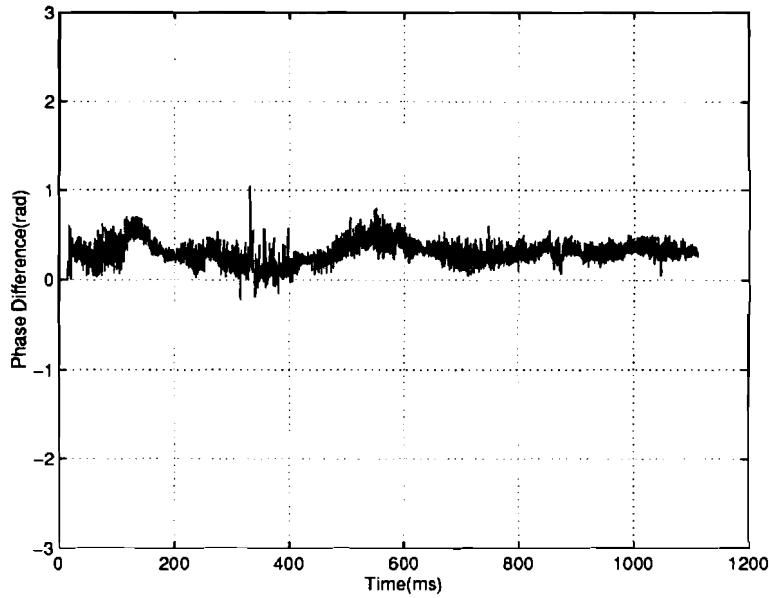


(a) Decoded $\hat{a}_1(k)$

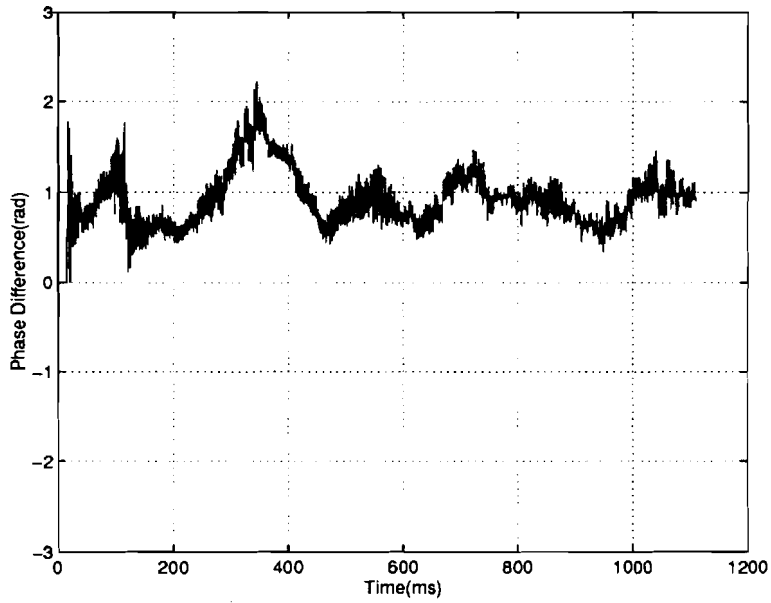


(b) Decoded $\hat{a}_2(k)$

Fig. 5.14. The CELP decoded $\hat{a}_i(k)$ for "Where were you while we were away"



(a) Decoded $\hat{\delta}_1(k)$



(b) Decoded $\hat{\delta}_2(k)$

Fig. 5.15. The CELP decoded $\hat{\delta}_i(k)$ for "Where were you while we were away".

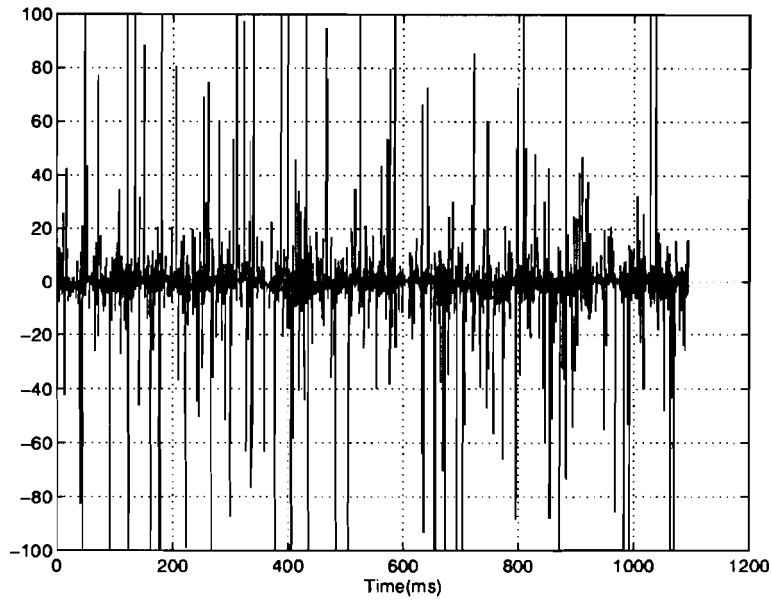
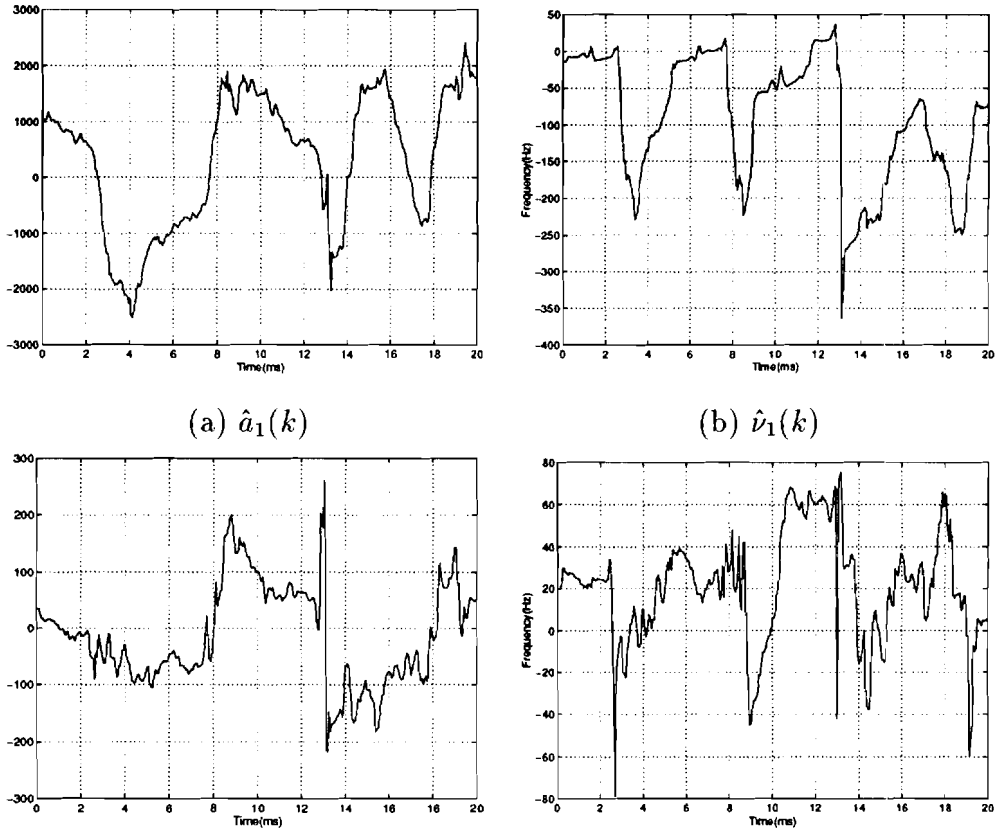


Fig. 5.16. The ratio $\frac{\cos(\hat{\phi}_1(k))}{\cos(\check{\phi}_1(k))}$ for "Where were you while we were away"

than attributing all of the variations to the signals being estimated. Therefore, by increasing r, we attempt to make a trade-off between the quality of the reconstructed speech and the suitability for coding of the MBDA estimates. Figure 5.17 shows the resulted estimates for r = 10 for the phoneme /ee/ of the word m/ee/ting processed in previous chapters (Figure 3.9). (The rest of the parameters remain the same.) These estimates appear to be "smoother" than they used to be. But the improvement is not substantial and the pitch synchronous behavior is not strengthened.

In our real speech experiments, the estimated phase signals appear to be quasi-linear or piecewise linear, e.g., the results displayed in Figure 3.11. Thus a very simple coding idea is to code and transmit every L th sample of $\phi_i(k)$ and then, at the receiver, recover the missing samples through linear interpolation. Figure 5.18 shows the error signal for the linearly interpolated $\hat{\phi}_1(k)$ for the sentence "Where were you while we were away" when L = 30 samples. Unfortunately, the error is noise like and is not pitch synchronous a.s we would like to observe. In addition the



(a) $\hat{a}_1(k)$ (b) $\hat{\nu}_1(k)$

(c) $\hat{a}_2(k)$ (d) $\hat{\nu}_2(k)$

Fig. 5.17. EKF estimates for the phoneme /ee/ of the word m/ee/ting: $r = 10$.

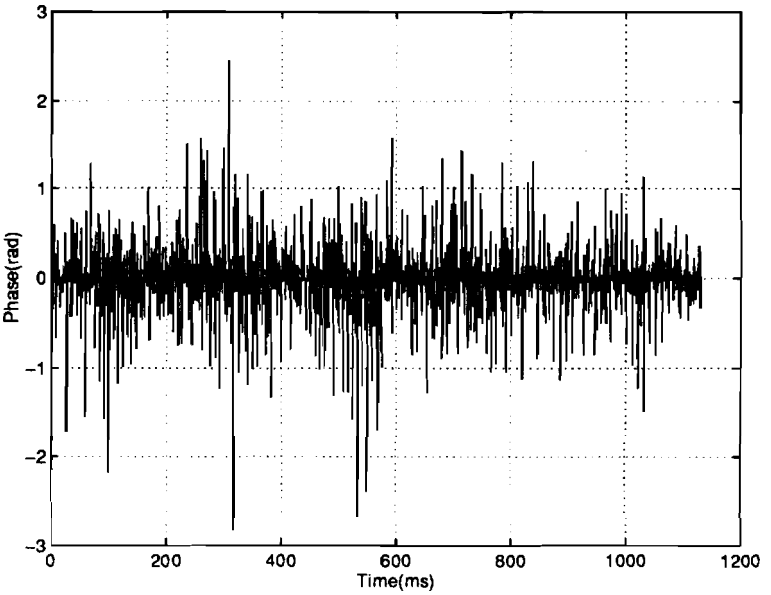


Fig. 5.18. The error for linearly interpolated $\hat{\phi}_1(k)$: L = 30.

error is of large amplitude relative to π. It would require a fair amount of effort to code this signal.

We have also experimented with a baseband coding approach. The idea is to shift the term $\hat{a}_2(k)\cos(\hat{\phi}_2(k))$ to baseband to generate an inphase signal and quadrature signal. Based on the inphase and quadrature signals, we can compute the envelope and phase which are subsequently coded (Figure 5.19). We are interested in this approach because LPC- or CELP-based schemes seem to be much less effective on estimates from the second resonance. If $\hat{a}_2(k)\cos(\hat{\phi}_2(k))$ is well behaved at baseband, it is then possible to code its envelope and phase efficiently. Figure 5.20 shows the estimated second resonance of the phoneme /ee/ of the word m/ee/ting in the time and frequency domains. The lowpass filter is shown in Figure 5.21 and the resulted envelope and phase signals are give11in Figure 5.22. These signals are observed to be very erratic and the phase signal does not exhibit strong pitch synchronous pattern. Therefore, this approach was not pursued.

So far, we have encountered substantial difficulties in coding MRDA outputs.

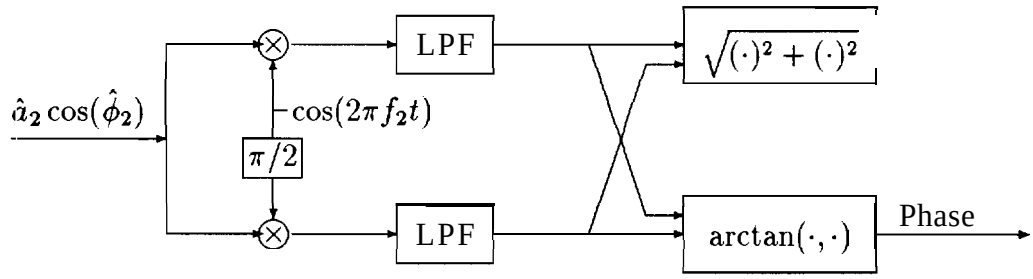


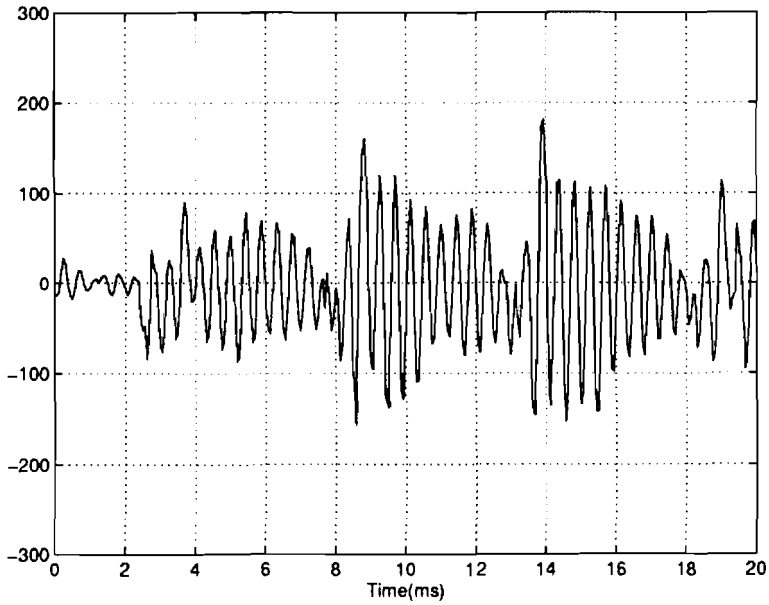
Fig. 5.19. The blockdiagram of baseband coding.

In fact, after the AM-FM model and its demodulation algorithms emerged, the application of the model to speech coding has been a very active research area [12]. However, no major breakthrough has been reported. We believe the difficulties come from several aspects: 1) The model is a nonlinear model. Our main efforts have been on coding amplitudes and phases. It is very difficult to control the effect of the phase signal on speech; and 2) The coding schemes that our experiments are based upon, i.e., LPC and CELP, are designed for speech signals rather than for MBDA estimates. This is manifested by the prestored excitation waveform in LPC-10 and the highly structured codebooks in CELP. This latter observation prompted us to search for a coding scheme that is not speech specific, e.g., subband coding.

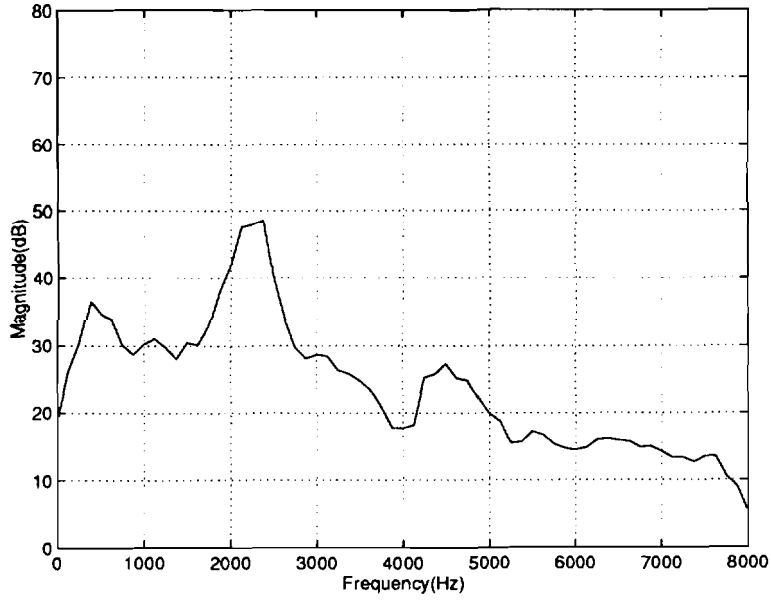
5.5 **Subband** Coding Approach

Generally in subband coding [42, 43, 44], a signal is passed through a bank of bandpass filters. The output of each filter is then coded and transmitted. Decimation is often involved. The idea of combining MBDA and subband coding approach for speech coding is shown in Figure 5.23. A speech signal is first passed through a bank of analysis filters. MBDA with one formant and a fixed formant frequency, i.e., $f_i(k+1) = f_i(k)$, is applied to the output of each filter to generate an amplitude and a phase for each band. These signals are then down-sampled, coded, and transmitted. A reverse process is performed at the receiver.

The approach is interesting because, in theory, the bandwidth of $\hat{a}_1(k)$ and $\hat{\delta}_1(k)$



(a) Time domain



(b) Frequency domain

Fig. 5.20. The estimated second resonance of the phoneme /ee/ of the word m/ee/ting.

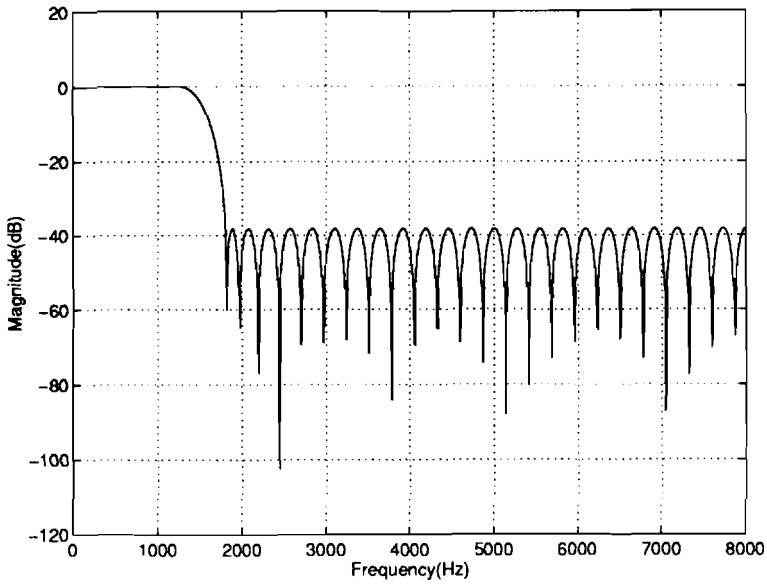
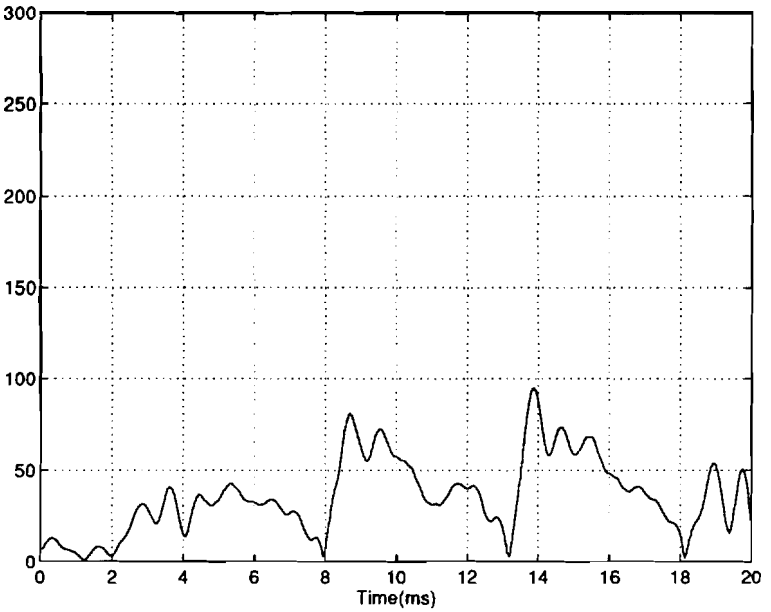


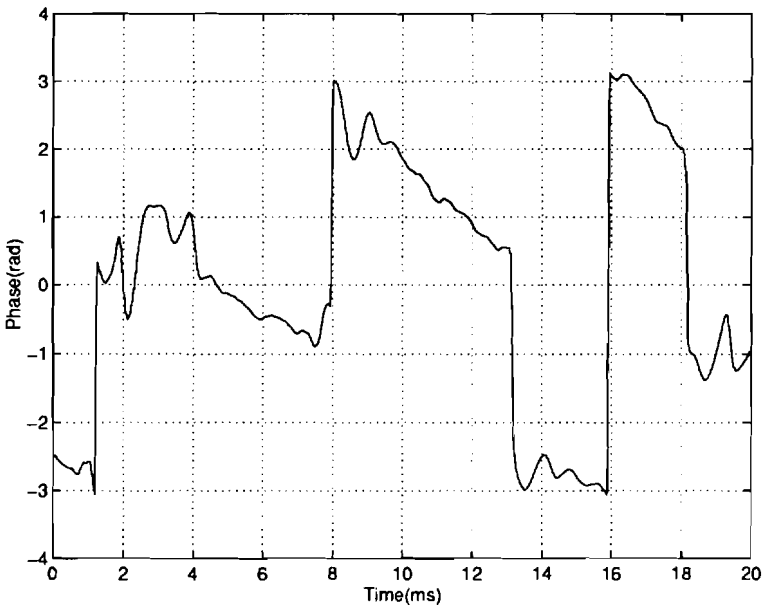
Fig. 5.21. The lowpass filter in baseband coding.

(or $\hat{\nu}_1(k)$) is much narrower than that of $y(k)$ where $y(k) = \hat{a}_1(k)\cos(\hat{\phi}_1(k))$ [29]. In other words, the nonlinear system exhibits bandwidth expansion analogous to the bandwidth expansion of a frequency-modulated communication system. In order to easily demonstrate bandwidth expansion, it is convenient to have a wide-sense stationary R_y (Eqs. (2.1)–(2.5)) and therefore we set $q_{f_i} = 0$. Unfortunately, we are unable to analytically compute $S_y(\Omega)$. Therefore, we give an numerical example using one formant and evaluating $S_y(\Omega)$ numerically after truncating the infinite sum to $\sum_{k=-4096}^{+4096}$. The parameters of the example are $a_{,,} = .99$, $q_{a_1} = 1$, $a_{,,} = .99$, $q_{\nu_1} = 20$, $q_{f_1} = 0$, $r = 0$, $m_{f_1,0} = 1000$ Hz, T = 1/16000 s, $p_{f_1,0} = 0$, and $p_{\phi_1,0} = 0$. In Figure 5.24 we show the power spectral densities for $a_1(k)$ and ν_1(k), which are identical except for a constant scaling by 20^2, and the power spectral density for $y(k)$. Depending on the details of the definition of bandwidth, a bandwidth expansion of roughly **3** times has taken place in this example. Thus, this MBDA-subband coding idea is promising since down-sampling of a large rate could be achieved.

The coding method based on subband ideas is quite different from that based on



(a) Envelope



(b) Phase

Fig. 5.22. The envelope and phase at baseband for the phoneme /ee/ of the word m/ee/ting.

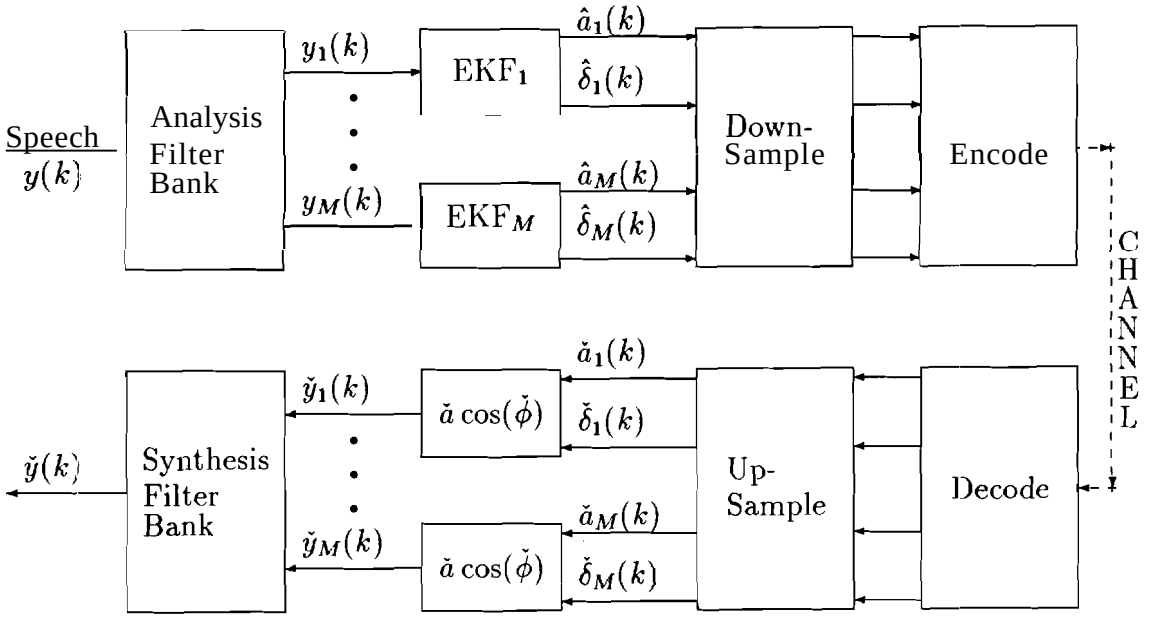


Fig. 5.23. The blockdiagram of MBDA–subband coding

LPC or CELP. In modified LPC or CELP, MBDA is applied to the entire speech. In the subband-based approach, MBDA is applied to the bandpassed speech with one formant and a fixed formant frequency. Both LPC and CELP assume a underlying linear prediction model while no model is assumed in subband coding.

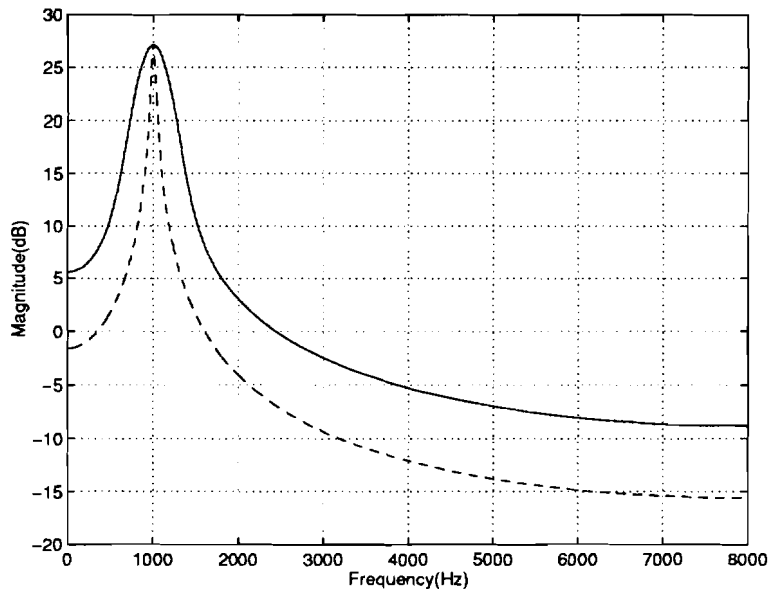


Fig. 5.24. Bandwidth Expansion: solid curve is $S_y(\Omega)$; dashed curve is the power spectral density of $a_1(k)$ shifted in frequency to $m_{f_1,0}$ and scaled in amplitude to match $S_y(\Omega)$, i.e., $(S_y(m_{f_1,0})/S_{a_1'}(m_{f_1,0}))S_{a_1'}(\Omega)$ where $a_1'(k) = a_1(k)\cos(2\pi m_{f_1,0}kT)$. $S_{\nu_1}(\mathrm{R})$ is proportional to $S_{a_1}(\Omega)$.

6. DISCUSSION

In this study we propose a statistical approach to the decomposition of a signal into sums of jointly amplitude- and frequency-modulated subsignals. More specifically, we propose a model and an estimation problem, we compute the Cramer-Rao bound for the estimation problem, and we propose and demonstrate a practical nonlinear estimator. This framework is then applied to speech problems.

This approach has several attractive features: (1) The approach can analyze signals containing multiple formants directly without first having to decompose into subsignals each containing zero or one formant. (2) The presence of interfering signals (i.e., "noise") is naturally included in the approach, both in the model and in the processing where, for instance, no derivatives or first differences of the noisy signal are required. (3) The target signals are precisely described. For instance, the bandwidths of the AM and FM subsignals can be independently controlled and the rate of change of the formant frequency can be controlled or the current formant-frequency model can be replaced by a more sophisticated model, e.g., a model which enforces spline-like smoothness constraints on the formant frequency. Specification of the target signals then implies the structure and parameters of the nonlinear filter. Furthermore, this level of control in the specification of the target signals allows the incorporation of additional acoustical knowledge as such knowledge becomes available. (4) The statistical framework allows the computation of bounds on the performance of an optimal estimator. For instance, in this study we compute the Cramer-Rao bound which is a lower bound on the mean square error between the unknown signals and their estimates and therefore is an upper

bound on the performance of the estimator. Using such bounds, the performance of practical suboptimal estimators can be compared against an absolute standard of performance.

The application of this approach to speech coding is also discussed. The idea is to code and transmit the amplitude and phase signals generated by our nonlinear filtering methods. We have experimented with a variety of techniques to code these estimated signals. It is shown that when standard linear prediction-based techniques are adopted, the advantage of this approach is that the necessary linear prediction model order is dramatically reduced so that fewer coefficients need to be transmitted. However, coding the residuals of the lineaa predictor is not straightforward since the methods embedded in standard speech coders (e.g., FS-101.5 and FS-1016) are specialized for speech signals (e.g., the prestored excitation sequence in FS-1015 and the small and highly structured codebooks in FS-1016). We believe that a non-speech specific technique, such as subband coding, will generate better results because no underlying linear prediction model is assumed and the demodulation is through MBDA with a fixed formant frequency where less ambiguity can occur.

BIBLIOGRAPHY

[1] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In W. J. Hardcastle and A. Marchal, editors, Speech Production and Speech Modelling, pages 241–261. NATO Advanced Study Institute Series D, vol. 55, 1990.

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

[3] 'Thomas Parsons. Voice and Speech Processing. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[4] 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.

[5] 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.

[6] B. van der Pol. Frequency modulation. Proc. IRE, vol. 18:pp. 1194–1205, July 1930.

[7] B. van der Pol. The fundamental principles of frequency modulation. IEE *J(London)*, vol. 93:pp. 153–158, 1946.

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

[9] Krishna S. Nathan, Yi-Teh Lee, and Harvey F. Silverman. A time-varying analysis method for rapid transitions in speech. IEEE *Trans.* Sig. *Proc.*, 39(4):815–824, April 1991.

[10] Krishna S. Nathan and Harvey F. Silverman. Time-varying feature selection and classification of unvoiced stop consonants. IEEE Trans. Speech Audio Proc., 2(3):395–405, July 1994.

[11] J. T. Foote, D. J. Mashao, and H. F. Silverman. Stop classification using DESA-1 high-resolution formant tracking. In Proceedings: 1993 International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 720–723, Minneapolis, MN, April 27–30 1993. IEEE–Signal Processing Society.

[12] 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.

[13] M.R. Schroeder. Vocoders: analysis and synthesis of speech. Proc. IEEE, 54:720–734, May 1966.

[14] J. L. Flanagan and R. M. Golden. Phase vocoder. Bell *System* Technical Journal, 45:1493–1509, November 1966.

[15] R. J. McAulay and T. F. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In Proc. IEEE *ICASSP-84*, pages 27.6.1–27.6.4, 1984.

[16] Robert J. McAulay and Thomas F. Quatieri. Mid-rate coding based on a. sinusoidal representation of speech. In Proc. IEEE *ICASSP-85*, pages 25.3.1–25.3.4, 1985.

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

[18] 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.

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

[20] Harry L. van Trees. Detection, Estimation, and Modulation *Theory:* Part I. John Wiley and Sons, Inc., New York, 1968.

[21] Harold W. Sorenson. Parameter Estimation: Principles and Problems. Marcel Dekker, New York, 1980.

[22] Louis L. Scharf. Statistical Signal Processing: Detection, Estimation, and Time Series Analysis. Addison-Wesley, Reading, MA, 1991.

[23] H. Vincent Poor. An *Introduction* to Signal Detection and Estimation. Springer-Verlag, New York, 1988.

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

[25] Jorge I. Galdos. A Cramer-Rao bound for multidimensional discrete-time dynamical systems. *IEEE Trans. Auto. Contr.*, 25(1):117–119, February 1980.

[26] Ben Zion Bobrovsky and Moshe Zakai. A lower bound on the estimation error for Markov processes. *IEEE Trans. Auto. Contr.*, 20(6):785–788, December 1975.

[27] Peter C. Doerschuk. Cramer-Rao bounds for discrete-time nonlinear filtering problems. *IEEE Trans. Auto. Contr.*, 40(8):1465–1469, August 1995.

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

[29] Shan Lu and Peter C. Doerschuk. Time-frequency analysis using dynamic statistical models with applications to speech analysis. In *Proceedings: 33 rd Allerton Conference on Communication, Control, and Computing.* Univ. of Illinois at Urbana-Champaign, October 4–6 1995. Accepted.

[30] Shan Lu and Peter C. Doerschuk. Demodulators for AM-FM models of speech signals: A comparison. In *Proceedings: 1996 International Conference on Acoustics, Speech, and Signal Processing,* Atlanta, Georgia, May 7–10 1996. IEEE–Signal Processing Society. Submitted.

[31] Shan Lu and Peter C. Doerschuk. Nonlinear modeling and processing of speech based on sums of AM-FM formant models. *IEEE Trails. Sig. Proc.* To appear as a regular paper.

[32] Shan Lu and Peter C. Doerschuk. Modeling and processing speech with sums of AM-FM formant models. In *Proceedings: 1995 International Conference on Acoustics, Speech, and Signal Processing,* pages 764–767, Detroit, MI, May 8–12 1995. IEEE–Signal Processing Society.

[33] National Communication System—Officeof Technology and Standards. *Federal Standard 1015, Telecommunications: Analog to digital conversion of radio voice by 2400 bit/second linear predictive coding, national communication system,* November 1984.

[34] National Communication System—Officeof Technology and Standards. *Federal Standard 1016, Telecommuizications: Analog to digital conversion of radio voice by 4800 bit/second code excited linear prediction (CELP),* February 1991.

[35] Thomas E. Tremain. The government standard linear predictive coding algorithm: LPC-10. *Speech Technology,* pages 40–49, April 1982.

[36] Panos E. Papamichalis. *Practical Approaches To Speech Coding.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[37] National Communication System—Office of Manager. Details *to* assist in implementation of federal standard 1016 CELP, January 1992.

[38] Joseph P. Campbell, Jr., Thomas E. Tremain, and Vanoy C. Welch. The proposed federal standard 1016 4800 bps voice coder: CELP. Speech Technoly, pages 58–64, April/May 1990.

[39] Manfred R. Schroeder and Bishnu S. Atal. Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In Proc. IEEE ICASSP-85, pages 937–940, 1985.

[40] Juin-Hwey Chen, Richard V. Cox, Yen-Chun Lin, and etc. A low-delay CELP coder for the CCITT 16 kb/s speech coding standard. IEEE *Journal* on Selected Areas in Communications, 10(5):830–849, June 1992.

[41] Ira A. Gerson and Mark A. Jasiuk. Vector sum excited linear prediction (VSELP) speech coding at 8kbps. In *Proc.* IEEE ICASSP-90, pages 461–464, 1990.

[42] R. E. Crochiere. On the design of subband coders for low bit-rate speech communications. Bell System *Technical Journal*, 56:747–770, May-June 1977.

[43] D. Esteban and Galand C. Application of quadrature mirror filters to split band voice coding schemes. In Proc. IEEE ICASSP-77, pages 191–195, 1977.

[44] Richard V. Cox, Steven L. Gay, Yair. Shoham, and etc. New directions in subband coding. IEEE Journal *on* Selected Areas in *Communications*, 6(2):391–409, February 1988.

[45] Jorge I. Galdos. A lower hound on filtering error with application to phase demodulation. IEEE Trans. Info. Theory, 25(4):452–462, July 1979.

[46] R. S. Bucy and P. D. Joseph. *Filtering* for Stochastic Processes With Applications to Guidance. Wiley, New York, 1968.

[47] R. E. Mortensen. Optimal control of continuous time stochastic systems. PhD thesis, Univ. of California., Berkeley, CA, 1966.

[48] T. E. Duncan. Probability densities for diffusion processes. PhD thesis, Stanford Univ., Stanford, CA, 1967.

[49] Shan Lu and Peter C. Doerschuk. Performance bounds for nonlinear filters. IEEE Trans. Aero. Elect. Syst. In review.

[50] Andrew J. Viterhi. Principles *of* Coherent Communication. McGraw-Hill Book Company, New York, 1966.

[51] Paul Bratley, Bennett L. Fox, and Linus E. Schrage. A Guide to Simulation. Springer-Verlag, 1983.

A. INITIAL CONDITIONS FOR EQ. (2.25)

In this appendix we describe the values for m^0 and Λ^0 which are the initial conditions for Eq. (2.25). Define $\mu_{f_i} \doteq (2\pi T m_{f_i,0}, 0)^T$, $\mu_{\nu_i} \doteq (0,0)^T$, $\mu_{a_i} \doteq (0,0)^T$, $= p^2_{\phi_i,0}/2$, $\rho = q^2_{\nu_i}/(1-\alpha^2_{\nu_i})$, $\eta = q^2_{a_i}/(1-\alpha^2_{a_i})$,

$$\Delta_{f_i} \doteq \begin{bmatrix} (2\pi T)^2 p^2_{f_i,0} + \kappa & \kappa \\ \kappa & \kappa \end{bmatrix},$$

$$\Delta_{\nu_i} \doteq \begin{bmatrix} (2\pi T)^2 \rho + \kappa & \kappa \\ \kappa & \kappa \end{bmatrix},$$

$$\Delta_{a_i} \doteq \begin{bmatrix} \eta & \alpha_{a_i}\eta \\ \alpha_{a_i}\eta & \eta \end{bmatrix},$$

$\mu_i \doteq (\mu_{a_i}, \mu_{f_i}, \mu_{\nu_i})^T$, $\Delta_i \doteq \text{diag}(\Delta_{a_i}, \text{A}\ ,\ ,\text{A},,)$, $\mu \doteq (\mu_1, \ldots, \mu_I)^T$, $\text{A} = \text{diag}(\Delta_1, \ldots . \Delta_I)$, and $\lambda \doteq (a_1(1), a_1(0),\ \phi_{f_1}(1), \phi_{f_1}(0),\ \phi_{\nu_1}(1), \phi_{\nu_1}(0),\ \ldots,\ a_I(1), a_I(0),\ \phi_{f_I}(1), \phi_{f_I}(0),\ \phi_{\nu_I}(1), \phi_{\nu_I}(0))^T$. Then $\lambda \sim \mathcal{N}(\mu, \Delta)$. Define $\xi \doteq (a_1(1), \phi_{f_1}(1), \phi_{\nu_1}(1),\ \ldots, a_I(1), \phi_{f_I}(1),\ \phi_{\nu_I}(1),\ a_1(0), \phi_{f_1}(0), \phi_{\nu_1}(0),\ \ldots,\ a_I(0), \phi_{f_I}(0), \phi_{\nu_I}(0))^T$ which is **a** permutation of X and define $\mathbf{P} \in \mathcal{R}^{6I \times 6I}$ which is the corresponding permutation matrix. **P** has values

$$P_{i,j} = \begin{cases} \delta_{i,(j+1)/2}, & 1 \le i \le 3I \\ \delta_{i-3I,j/2}, & 3I+1 \le i \le 6I \end{cases}$$

and satisfies $\xi = P\lambda$. Therefore, $m^0 = \mathbf{E}[\xi] = P\mu$ and $\Lambda^0 = \mathbf{E}[(\xi - \mathbf{E}[\xi])(\xi - \mathbf{E}[\xi])^T] = \text{P A P}^{\text{T}}$.

B. AN ALTERNATIVE PERFORMANCE BOUND

In this appendix, we describe an alternative to the Cramer-Rao bound, specifically, a lower bound on mean square error performance based on rate distortion theory. We also describe a Monte Carlo method for evaluating the bound.

We consider the following discrete time model [45]:

$$x_{k+1} = a(x_k, k) + b(x_k, k)w_k \tag{B.1}$$

$$y_k = g(x_k, k) + N(k)v_k \tag{B.2}$$

where $x_k \in \mathrm{R}^n$, $y_k \in \mathrm{R}^m$, w_k is i.i.d. $\mathcal{N}(0, I_n)$, v_k is i.i.d. $\mathcal{N}(0, \mathrm{I},)$, x_0 is $\mathcal{N}(\bar{x}_0, \Sigma_0)$, and w, v, and x_0 are all independent. Notation: $x_i^j = \{x_i, x_{i+1}, \ldots, x_j\}$ and likewise for y_i^j, v_i^j and w_i^j. $R(j) = N(j)N(j)^T$. E is expectation and E^{x*} is conditional expectation given x_k.

In general, the design of a filter that estimates x_k as a function of the observations $\{y_\sigma, \sigma \leq k\}$ is a nonlinear filtering problem. The goal is often to minimize the mean square error measure of distortion:

$$\varepsilon(k) = \mathbf{E}\{(x_k - \hat{x}_k)^T(x_k - \hat{x}_k)\},$$

where $\hat{x}_k$ is the estimate of x_k.

The optimal solution that minimizes the mean square error is

$$\hat{x}_k^* = \mathbf{E}\{x_k | y_\sigma, \sigma \leq k\}$$

with the associated optimal error

$$\varepsilon^*(k) = \mathbf{E}\{(x_k - \hat{x}_k^*)^T(x_k - \hat{x}_k^*)\}. \tag{B.3}$$

In practice, the optimal estimator can not be built, nor is it possible to compute the optimal error. However, there exist some lower bounds on on the mean square error which provide an indication of whether accuracy requirements are realistic before undertaking a suboptimal filter design. One such bound based on rate distortion theory and the Bucy-Mortensen-Duncan representation theorem [46, 47, 48] is originally suggested in [45] and subsequently corrected in [49]. The bound is given by Theorem 1.

Theorem *1 Consider the discrete time filtering problem. defined by Eqs. (B.1) and (B.2). A lower bound on the optimal mean square error* $\varepsilon^*(k)$ *of Eq. (B.3) is given by*

$$\varepsilon^*(k) \geq \frac{n}{2\pi e} \exp\left\{\frac{2}{n} h(p_{x_k})\right\} \cdot \exp\left\{-\frac{2}{n} I(x_k; y_0^k)\right\}, \tag{B.4}$$

where $h(p_x)$ *denotes the differential entropy of a random vector* x *with probability density function (pdf)* p_x*, i.e.,*

$$h(p_x) = -\int \log[p_x(\eta)] p_x(\eta) \mathrm{d}\eta,$$

$$\begin{aligned} I(x_k; y_0^k) = \mathbf{E}\Big\{ & \log\left[\mathbf{E}^{x_k}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k = y_0^k} \\ & - \log\left[\mathbf{E}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k = y_0^k} \Big\} \end{aligned} \tag{B.5}$$

where z_0^k *is a deterministic vector of the same dimensions as* y_0^k*, and*

$$\begin{aligned} \zeta_t(x_0^k, y_0^k) &= \sum_{j=0}^{k} g(x_j, j)^T R^{-1}(j) y_j \\ &\quad - \frac{1}{2} \sum_{j=0}^{k} g(x_j, j)^T R^{-1}(j) g(x_j, j) \\ &= \frac{1}{2} \sum_{j=0}^{k} g(x_j, j)^T R^{-1}(j) g(x_j, j) \\ &\quad + \sum_{j=0}^{k} g(x_j, j)^T R^{-1}(j) N(j) v_j(x_j, y_j). \end{aligned} \tag{B.6}$$

Proof: See Appendix C. □

The difficulty with applying Theorem 1 to a particular problem is that Eq. (B.5) involves conditional expectations in function space. For a limited class of problems, we now describe a Monte Carlo method for evaluating Eq. (B.5) and consequently the bound in Eq. (B.4). Specifically, we describe a method for sampling from the conditional distribution (in the Gaussian case) and an importance sampling method which accelerates the convergence of the Monte Carlo sums. Though the Gaussian restriction is limiting, there exist applications, such as analog angle modulation systems [50] and speech processing problems of this report [32], where the state process is linear and Gaussian but the measurement equation is nonlinear, and these methods are oriented toward such problems.

Any practical sampling algorithm must operate by transforming a set of i.i.d. $\mathcal{N}(0,1)$ samples. The sampling method we propose is based on the following observation. If x is $\mathcal{N}(\bar{x}, \mathrm{C})$ and x is partitioned $x = (x_a^T, x_b^T)^T$ (and likewise for $\bar{x}$ and Σ) then [28, p. 321] $p_{x_a|x_b}(x_a|x_b) = \mathcal{N}(\bar{x}_{a|b}(x_b), \Sigma_{a|b})(x_a)$ where $\bar{x}_{a|b}(x_b) = \bar{x}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \bar{x}_b)$ and $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab}^T$. The conditional mean estimator of x_a based on x_b, denoted by $\hat{x}_{a|b}$, is $\hat{x}_{a|b} = \bar{x}_{a|b}$ and the error, denoted by $\tilde{x}_{a|b}$, is defined to be $\tilde{x}_{a|b} = x_a - \hat{x}_{a|b}$ and has the properties that $\mathbf{E}\{\tilde{x}_{a|b}\} = 0$ and $\mathrm{Var}\{\tilde{x}_{a|b}\} = \Sigma_{a|b}$. Therefore, samples can be drawn from $\mathcal{N}(\bar{x}_{a|b}(x_b), \Sigma_{a|b})$ for a particular value of x_b (denoted by x_b^*) by using the following algorithm:

1. Choose x_b^*.

2. Pick a realization of x.

3. Compute $\hat{x}_{a|b}$.

4. Compute $\tilde{x}_{a|b}$. $\tilde{x}_{a|b}$ is distributed $\mathcal{N}(0, \Sigma_{a|b})$.

5. Compute $m_a = \bar{x}_{a|b}(x_b^*)$. (This would actually be done only once during an initialization procedure).

6. Compute $z = \tilde{x}_{a|b} + m_a$. z is distributed $\mathcal{N}(\bar{x}_{a|b}(x_b^*), \Sigma_{a|b})$ as desired.

In the case where x is the sequence of states of a dynamical system (specifically, x_a is the first M time-steps and x_b is the M+1 st time-step), this process can be done efficiently because

1. The process noise w is a white sequence and therefore is easy to sample from, and the realization of x can easily be computed from the realization of w.

2. The conditional mean $\hat{x}_{a|b}$ can easily be computed by a Kalman smoother once an observation equation, with measurement matrix H and measurement noise covariance matrix R, is defined by

$$H_k = \begin{cases} 0, & k = 0, 1, \ldots, M-1 \\ I, & k = M \end{cases}$$
$$R_k = \begin{cases} I, & k = 0, 1, \ldots, M-1 \\ 0, & k = M \end{cases}$$

   A forward-backward two-filter algorithm [28, p. 189] is attractive because the forward pass is very simple.

Let χ be a random variable with probability density function p. Computation of $\mathbf{I} = \mathbf{E}[f(\chi)]$ using Monte Carlo and importance sampling [51, Section 2.5] requires the computation of $\hat{I}_q = (1/L)\sum_{i=0}^{L-1} f(\gamma_i)p(\gamma_i)/q(\gamma_i)$ where γ_i are i.i.d. samples from the probability density function q. We apply these ideas to the computation of the conditional expectation in Eq. (B.5). (The complete computation of the first term of Eq. (lB.5) requires Monte Carlo for the outer expectation also). Though the optimal q is known, we use a simple sub-optimal choice: p is Gaussian and we choose q to be Gaussian with the same covariance but a different mean. The new mean is chosen as a compromise between the x_0^k trajectory that maximizes $\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)$ (which corresponds to f in $\mathbf{I} = \mathrm{E}[\mathrm{f}(\chi)]$ above) and the mean of p, which maximizes p. The compromise is a time-step by time-step convex combination of the two trajectories, which are denoted m_ζ and m_p respectively. The weight in the convex combination at step i on the mean of p is $\sigma_\zeta^2/(\sigma_\zeta^2 + \sigma_p^2)$ where σ_ζ is the width, as a function of x_i, of the maximum of $\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)$ at the trajectory m_ζ and σ_p^2 is the i th

diagonal element of the covariance of p. The width is defined as the value of δx_i such that $\zeta_k\left(m_\zeta, z_0^k\right) - \zeta_k\left(\delta m_\zeta, z_0^k\right) = 2$ where $(\delta m_\zeta)_k = (m_\zeta)_k$ fur k $\neq i$ and $(\delta m_\zeta)_k = (m_\zeta)_k + \delta x_i$ for k $= i$ [$(\cdot)_k$ means the value at time step k]. Because of the form of ζ, which is due to the fact that the observation noise is white, it is easy to find the trajectory x_0^k that maximizes exp(ζ_k) because it can be done time-step by time-step.

In order to demonstrate these ideas, we have examined a scalar linear-Gaussian example: $x_{k+1} = ax_k + bw_k$ and $y_k = gx_k + Nv_k$ where a = 0.5, b = 0.5, g = 2, $N = 1$, and x_0 is $\mathcal{N}(0, 20)$. The exact mean-square error at k = 50, computed using the Kalman Filter, is $\Sigma_{50|50} = 0.1328$. Pencil and paper evaluation of the rate distortion bound (i.e., Eq. (B.5)) Ref. [45]) gives the same result. The Monte Carlo value based on L = 1000 terms in each inner integral and in the outer integral and without using importance sampling is 0.0438 ± 0.0358 which is clearly far from convergence. (The sample mean of the bound plus/minus 1 sample standard deviation of the bound based on 10 runs is reported). Finally, if the calculation is unaltered except for the use of importance sampling, then the value is 0.1044 ± 0.0370 which is substantially closer to the exact value.

Overall, compared to the Cramer-Rao hound described in Section 2.3, the bound basecl on the rate distortion theory is computationally more burdensome and therefore was not pursued.

C. PROOF OF THEOREM 1

The bound proposed in [45] is more attractive than the bound of Theorem 1 in Appendix B because in [45] there are no conditional expectations in function space. However, as described in [49], there is an error in the proof in [45], which invalidates the bound. While the original bound is incorrect, an intermediate result is correct and that result is Theorem 1. Therefore, we only describe the error in [45] that invalidates the more attractive but incorrect bound.

In [45], Theorems 2 (continuous time) and 3 (discrete time) have parallel proofs. An error occurs in the final step of the common proof between Eqs. (A.9) and (A10) in Appendix A and is of the following general type. Let r and s be scalar random variables on the same probability space and let $z \in \mathrm{R}$ be a deterministic parameter. Let h be a scalar random variable derived from r and s. Define two functions $f(s) = (\mathbf{E}\{h(r,z)\})|_{z=s}$ and $g(s) = \mathbf{E}\{h(r,s)|s\}$. In general, $f(s) \neq g(s)$. Furthermore., while $\mathbf{E}\{h(r,\mathrm{s})) = \mathbf{E}\{\mathbf{E}\{h(r,s)|s\}\} = \mathbf{E}\{g(s)\}$, in general it is not true that $\mathbf{E}\{h(r,\mathrm{s}))$ equals $\mathbf{E}\{\mathrm{f}(\mathrm{s}))$. The error leading up to Eq. (A10) is an assertion of the type that $\mathbf{E}\{h(r,\mathrm{s})) = \mathbf{E}\{\mathrm{f}(\mathrm{s}))$. A concrete example follows:

Let $r \in \{0,1\}$, $s \in \{0,1\}$, $p_{r,s}(0,0) = p_{0,0}$, $p_{r,s}(0,1) = p_{0,1}$, $p_{r,s}(1,0) = p_{1,0}$, $p_{r,s}(1,1) = 1 - p_{0,0} - p_{0,1} - p_{1,0}$, and $h(r,\mathrm{s}) = rs$. Then $\mathrm{f}(\mathrm{s}) = (\mathbf{E}\{h(r,z)\})|_{z=s} = s(1 - p_{0,0} - p_{0,1})$ and $g(s) = \mathbf{E}\{h(r,s)|s\} = s(1 - p_{0,0} - p_{0,1} - p_{1,0})/(1 - p_{0,0} - p_{1,0})$. Therefore, $\mathbf{E}[g(s)] = 1 - p_{0,0} - p_{0,1} - p_{1,0} = \mathbf{E}[h(r,s)]$ and $\mathbf{E}[\mathrm{f}(s)] = (1 - p_{0,0} - p_{0,1})(1 - p_{0,0} - p_{1,0})$ so that $\mathbf{E}[f(s)] \neq \mathbf{E}[g(s)] = \mathbf{E}[h(r,s)]$.

We focus on the discrete-time case where for each equation we give both the abstract form in terms of expectations and the concrete form in terms of integrals

and probability density functions (which we assume exist). We consider the same model given in Eqs. (B.1) and (B.2).

The Bucy-Mortensen-Duncan representation theorem in discrete time is

$$p(x_k|y_0^k) = \frac{\left[\mathbf{E}^{x_k}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k=y_0^k} p(x_k)}{\left[\mathbf{E}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k=y_0^k}}$$

where z_0^k is a deterministic vector of the same dimensions as y_0^k and $\zeta_k(x_0^k, y_0^k)$ is given in Eq. (B.6). (There is a typographical error in Eq. (26) of Ref. [45]: the "$N(j)$" factor in Eq. (B.6) is missing). More explicitly, the Bucy-Mortensen-Duncan representation theorem is

$$p(x_k|y_0^k) = \frac{\left[\int_{x_0^{k-1}} p(x_0^{k-1}|x_k)\exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^{k-1}\right] p(x_k)}{\left[\int_{x_0^k} p(x_0^k)\exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^k\right]}$$

Eq. (A10) of Ref. [45] is a bound on the mutual information $I(x_k; y_0^k)$ based on the Bucy-Mortensen-Duncan representation theorem. The derivation is correct through

$$\begin{aligned} I(x_k; y_0^k) &= \mathbf{E}\Big\{\log\left[\mathbf{E}^{x_k}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k=y_0^k} \\ &\quad - \log\left[\mathbf{E}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k=y_0^k}\Big\} \qquad \text{(C.1)} \\ &\leq \log\mathbf{E}\left\{\left[\mathbf{E}^{x_k}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k=y_0^k}\right\} \\ &\quad - \mathbf{E}\left\{\log\left[\mathbf{E}\left\{\exp\left(\zeta_k\left(x_0^k, z_0^k\right)\right)\right\}\right]_{z_0^k=y_0^k}\right\}. \qquad \text{(C.2)} \end{aligned}$$

(The second step is Jensen's inequality). The final step leading to Eq. (A10) of Ref. [45] is the assertion that $\mathbf{E}\{[\mathbf{E}^{x_k}\{\exp(\zeta_k(x_0^k, z_0^k))\}]_{z_0^k=y_0^k}\}$ equals $\mathbf{E}\{\exp(\zeta_k(x_0^k, y_0^k))\}$ which, unlike $\mathbf{E}\{\mathbf{E}^{x_k}\{\exp(\zeta_k(x_0^k, y_0^k))\}\} = \mathbf{E}\{\exp(\zeta_k(x_0^k, y_0^k))\}$, is incorrect. More explicitly, the bound is

$$\begin{aligned} & I(x_k; y_0^k) \\ = & \int_{x_k, y_0^k} p(x_k, y_0^k)\log\left\{\int_{x_0^{k-1}} p(x_0^{k-1}|x_k)\exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^{k-1}\right\} \mathrm{d}x_k \mathrm{d}y_0^k \\ & - \int_{y_0^k} p(y_0^k)\left\{\log\int_{x_0^k} p(x_0^k)\exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^k\right\} \mathrm{d}y_0^k \end{aligned}$$

$$\leq \log \int_{x_0^k, y_0^k} p(x_k, y_0^k) p(x_0^{k-1} | x_k) \exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^k \mathrm{d}y_0^k$$
$$- \int_{y_0^k} p(y_0^k) \left\{ \log \int_{x_0^k} p(x_0^k) \exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^k \right\} \mathrm{d}y_0^k$$

and the incorrect assertion is that

$$\int_{x_0^k, y_0^k} p(x_k, y_0^k) p(x_0^{k-1} | x_k) \exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^k \mathrm{d}y_0^k$$

equals

$$\int_{x_0^k, y_0^k} p(x_0^k, y_0^k) \exp\left(\zeta_k\left(x_0^k, y_0^k\right)\right) \mathrm{d}x_0^k \mathrm{d}y_0^k.$$

The incorrect assertion is important because, if it were true, then these would be no conditional function space expectations which are more difficult to evaluate than unconditional expectations. In the absence of the assertion, **Eq.** (C.1), which is identical to **Eq.** (B.5) of Appendix B, provides a tighter bound than **Eq.** (C.2) and, at least in a Monte Carlo approach, requires essentially the same amount of computation.