Speech Processing Applications Using an AM-FM Modulation Model

A thesis presented

by

Alexandros Potamianos

to

The Division of Applied Sciences
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Engineering Sciences

Harvard University
Cambridge, Massachusetts

August 1995

Abstract

In this thesis, the AM–FM modulation speech model and multiband demodulation are applied to speech analysis and coding. The AM–FM model represents the speech signal as a sum of amplitude modulated (AM) and frequency modulated (FM) signals; each AM–FM signal models a single speech resonance (formant). The model is able to describe a wide range of nonlinear and time–varying phenomena during speech production. Multiband demodulation is the proposed speech analysis method in the context of the AM–FM model. A bank of Gabor filters is used to filter the speech signal and, then, a demodulation algorithm is applied on each band to obtain the amplitude envelope and instantaneous frequency signals. The energy separation algorithm (ESA) and the Hilbert transform approach are compared for signal and speech resonance demodulation, and the ESA is found to have better time–resolution and to be computationally more efficient. Next, we apply multiband demodulation analysis (MDA) to formant and pitch tracking. Using the amplitude envelope and instantaneous frequency signals short–time estimates are proposed for the formant frequency and the fundamental frequency. The merits of the estimates are evaluated and it is concluded that the amplitude weighted mean instantaneous frequency and the short–time phase slope perform best for formant and pitch estimation respectively. Finally, decision algorithms are provided for the formant and pitch contours. Both speech analysis algorithms provide very smooth and accurate estimates and have attractive time–domain parallel implementations. Next, we use time–varying MDA for a speech coding application. A time–varying Gabor filterbank extracts four formant bands from the signal and, then, each resonance is demodulated to amplitude envelope and instantaneous frequency signals. Efficient modeling and coding schemes are proposed for the information signals that exploit the correlation between the formant bands. Finally, speech is synthesized as the sum of the reconstructed formant bands. The AM–FM analysis–synthesis system produces speech of very natural quality. Currently, the vocoder operates in the 4.8–9.6 kbits/sec range. Future applications of these modeling/coding ideas include text–to–speech synthesis and speaker identification. Overall, the AM–FM modulation model and multiband demodulation analysis are a general nonlinear approach to speech processing with a wide range of successful applications.

Acknowledgments

First I wish to thank my advisor Prof. Petros Maragos for a challenging research area, his friendship and his continuous support. His academic brilliance has been a guide in my research efforts. His contributions in this research work have been very significant.

My thanks go to a person I was not fortunate to meet, Dr. Herbert Teager, a pioneer whose work is an almost inexhaustible tank of ideas. I also thank Dr. Thomas Quatieri and Dr. James Kaiser for the interaction, support, and kindness they showed throughout the course of my graduate years; their comments and suggestions, along with those of Prof. Roger Brockett, have greatly improved the quality of this dissertation. I also thank Dr. Helen Hanson and Prof. Alan Bovic for direct and indirect contributions to this work.

Although, my work at Bell Labs is not related with this thesis, I wish to thank my mentor there, Dr. Richard Rose. For the short period since I have known him he has been both an inspiration and a friend. I also thank Prof. George Karagiannis for his kindness and for introducing me to signal and speech processing.

Next, I thank my friends at Harvard University, especially, Mike Yang, George Michail, Aris Moustakas and everybody in the robotics lab. My gratitude is extended to the person who was for many years the heart of the Division of Applied Sciences, Pauline Mitchell.

My friends at Georgia Tech, Mike Macon, Mathieu Hans, Haluk Aydinoglu, Mohamed Ben Romhdane and everybody in the DSP lab have helped me both in work related and personal matters throughout my two-year stay there. I thank them for their support.

The gratitude for my family to whom this thesis is dedicated would be out of place if it was to be discussed in a few lines. They are part of me always.

Last but not least, I wish to thank my life companion Alexandra for her inspiration, patience and support.

Contents

Notation

$x(t)$, $x(n)$	General continuous- and discrete-time signals
$s(t)$, $s(n)$	Continuous- and discrete-time speech signals
$r(t)$, $r(n)$	Continuous- and discrete-time speech resonance signals
Γ, σ	Damping parameter for continuous- and discrete-time 2nd order systems
$\Psi_c[.]$, $\Psi_d[.]$, $\Psi[.]$	Continuous- and/or discrete-time energy operator
$\Upsilon_k[.]$	Continuous- or discrete-time kth-order energy operator
$a(t)$, $a(n)$	Continuous- and discrete-time amplitude envelope signal
$a_H(t)$, $a_H(n)$	Amplitude envelope signal estimated by the HTD
$a_E(t)$, $a_E(n)$	Amplitude envelope signal estimated by the ESA
$\omega(t)$, $\Omega(n)$	Continuous- and discrete-time angular instantaneous frequency signal
$\omega_H(t)$, $\Omega_H(n)$	Angular instantaneous frequency signal estimated by the HTD
$\omega_E(t)$, $\Omega_E(n)$	Angular instantaneous frequency signal estimated by the ESA
$f(t)$	Continuous-time instantaneous frequency signal
$f_H(t)$	Instantaneous frequency signal estimated by the HTD
$f_E(t)$	Instantaneous frequency signal estimated by the ESA
$\phi(t)$, $\phi(n)$	Continuous- and discrete-time phase signal
$\phi_H(t)$, $\phi_H(n)$	Phase signal estimated by the HTD
$\phi_E(t)$, $\phi_E(n)$	Phase signal estimated by the ESA
$q(t)$	Frequency modulating signal
ω_c, Ω_c	Angular carrier frequency for continuous and discrete AM–FM signals
f_c	Carrier frequency
ω_m, Ω_m	Maximum angular frequency deviation for AM–FM signals
ω_m/ω_c, Ω_m/Ω_c	Modulation depth for continuous and discrete AM–FM signals

κ	Amplitude modulation index
W_a	Bandwidth of the amplitude modulating signal
W_f	Bandwidth of the frequency modulating signal
$\Omega_c/W_{a,f}$	Carrier frequency to information bandwidth ratio
$z(t)$	Gabor's analytic signal
$h(t)$, $H(w)$	Impulse and frequency response of the Gabor filter
ν	Center frequency of the Gabor filter
F, F_1–F_4	Formant frequencies
B, B_1–B_4	Formant bandwidths
F_0	Fundamental frequency
F_u, F_w	Unweighted and weighted short-time formant frequency estimates
B_u, B_w	Unweighted and weighted short-time formant bandwidth estimates
S_ϕ	Short-time phase slope estimate
F_s	Sampling frequency
$\mathcal{N}$	Set of natural numbers
$\mathcal{Z}$	Set of integers
$\lfloor . \rfloor$, $\lfloor . + 0.5 \rfloor$	Decimal part truncation and rounding operator
$\mathcal{ES}(\nu)$	Energy spectrum

Chapter 1

Introduction

1.1 Speech Production

A broad definition for *sound* is the phenomena in air responsible for the sensation of hearing. Sound is produced when changes in air density and particle air velocity are translated into pressure variations.

At the human vocal system, sound is produced when air is pushed from the lungs through the *glottis* into the *vocal tract*. The air flow is turned into a pressure wave at the glottis or at a constriction along the vocal tract. The tract consists of the glottis, the pharynx and the oral cavity. For nasal sounds, the velum is lowered and the nasal tract becomes acoustically coupled to the vocal tract. The shape of the vocal tract is defined by the position of the articulators: the tongue, the lips, the jaw and the musculature of the pharynx. During speech production, the articulators move smoothly from one position to another to produce different sounds.

Each distinctive speech sound is called a *phoneme*; in the English language there are approximately 42 phonemes. Phonemes are grouped according to the primary excitation source and the configuration of the articulators during phonation. For *voiced* sounds the primary excitation is at the glottis: the vocal cords vibrate in a relaxation oscillation that produces quasi–periodic pressure pulses. The frequency and period of the oscillations of the vocal cords are called the *fundamental frequency* and the *pitch period* respectively. For *unvoiced* sounds the vocal cords are inactive and the glottis is open. The excitation

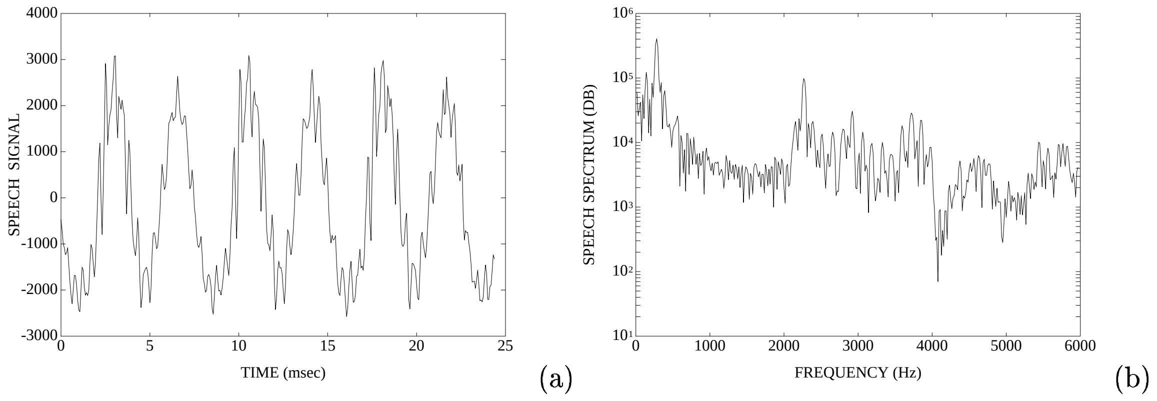


Figure 1.1: (a) Three pitch periods of the vowel /ih/ sampled at 16 kHz, (b) the short–time Fourier transform of the vowel.

for unvoiced *fricative* phonemes is located at a constriction at the larynx or along the vocal tract. The constriction causes high air velocities (high Reynolds number [106]) and turbulent air flow. The resulting pressure wave has a broad noise–like spectrum. Finally, *plosive* sounds are produced by an abrupt release of air stored behind a total closure in the front part of the vocal tract, the teeth or the lips. Note that secondary excitations also contribute to the speech waveform. For example, for voiced speech the excessive airflow produces *aspiration* noise at the glottis. More than one of the speech sources described above can be active at one time.

The vocal tract can be thought of as the cascade of a few (typically 4–5) vocal cavities. When exited by the source waveform each cavity produces an oscillatory response at the characteristic frequency of the cavity, the *resonant frequency.* Overall, the tract acts as a filter that amplifies the characteristic frequencies of the vocal cavities. As a result, the Fourier transform of a speech waveform has dominant spectral peaks (called *formants*) that correspond to the vocal tract resonances. In Fig. 1.1 (a), (b), we show three pitch periods of the vowel /ih/ and the Fourier transform of this 25 msecs speech segment. Note that the speech formants are evident in (b) as spectral peaks (approximately) at 300, 2200, 3000 and 3800 Hz. The pitch periodicity of the speech waveform is manifested in (b) by the harmonic structure of the Fourier spectrum. An overview of the fundamentals of speech production can be found in [18, 92].

1.2 Speech Analysis: The Linear Model

The equations of acoustics originate from the conservation of mass and conservation of momentum equations of fluid mechanics [106]. To model the acoustics of speech production these nonlinear equations are simplified to a one–dimensional linear form using assumptions that include the following [82, 92]:

Linear Model Assumptions

1. The air flow fills up the whole vocal tract area, and the air flow velocity profile is uniform across a cross–section of the vocal tract.

2. The air flow velocity is much smaller than the sound velocity.

3. The fluid cannot support shear stresses (no viscosity).

These assumptions allows us to discard the nonlinear terms in the Navier–Stokes equation. For a tube with a constant cross–sectional area A the linear acoustics equations relating the sound pressure $p(x,t)$ with the volume velocity flow $v(x,t)$ are

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial v}{\partial t} \qquad -\frac{\partial v}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{1.1}$$

where ρ is the *ambient* density of the air in the tube and c is the sound velocity. The solution of Eq. (1.1) is the superposition of waves traveling in the positive and negative directions [92]

$$v(x,t) = [v^+(t - x/c) - v^-(t + x/c)] \tag{1.2}$$

$$p(x,t) = \frac{\rho c}{A}\,[v^+(t - x/c) - v^-(t + x/c)] = \frac{\rho c}{A}\,v(x,t). \tag{1.3}$$

The important result here is that pressure and volume velocity are linearly related. From here it follows how to solve for v^+ and v^-. Correction terms for the vibrating walls effects, the effects of viscous friction, thermal conduction at the walls, and radiation from the lips have also been accounted for in the linear theory of speech production [92, 18].

The linear *source–filter* model assumes that the vocal tract is composed of typically 4–5 tubes/cavities of constant cross–sectional area. Each tube is modeled as a second–order damped linear resonator/filter (the damping accounts for losses due to friction). The impulse response of the filter is

$$r(n) = A\sigma^n \cos(2\pi(F/F_s)n + \theta), \qquad n > 0 \tag{1.4}$$

where F is the resonant frequency of the cavity (formant frequency), $\sigma \in [0,1]$ controls the energy dissipation rate (formant bandwidth parameter) and F_s is the sampling frequency. The corresponding transfer function is

$$R(z) \approx \frac{1 + b_1 + b_2}{1 + b_1 z^{-1} + b_2 z^{-2}} \tag{1.5}$$

where b_1 and b_2 are constants that can be expressed as functions of the formant frequency F and the bandwidth parameter σ. According to the linear theory, the vocal tract can be represented as a cascade of N linear resonators/filters with impulse response $w(n)$ and transfer function $W(z)$ given by:

$$w(n) = r_1(n) * r_2(n) * \cdots * r_N(n) \tag{1.6}$$

$$W(z) = \prod_{k=1}^{N} R_k(z) = \prod_{k=1}^{N} \frac{1 + b_{1k} + b_{2k}}{1 + b_{1k} z^{-1} + b_{2k} z^{-2}} = \frac{G}{1 - \sum_{i=1}^{2N} a_i z^{-i}} \tag{1.7}$$

where N is the number of speech resonances (formants) and the constant G controls the amplitude (gain). The source–filter model assumes that the glottis and the vocal tract are not coupled, i.e., the glottal pressure wave is assumed to excite the vocal tract in a similar way that a signal excites a linear filter. Thus, the speech source waveform $u(n)$ is convolved with the vocal tract response $w(n)$ to produce the speech waveform $s(n)$

$$s(n) = u(n) * w(n). \tag{1.8}$$

The block diagram of the source–filter linear model is shown in Fig. 1.2. According to the simplified excitation model of Fig. 1.2 the excitation $u(t)$ signal is either an impulse train or random noise for voiced and unvoiced speech respectively.

Speech is a non–stationary process, yet most of the speech features (e.g., pitch period, formant frequencies) are slow varying with time. For estimation purposes the parameters of the linear model are assumed to be constant over a short–time analysis window (typical analysis window duration is 10–30 msecs). A very popular mathematical framework for

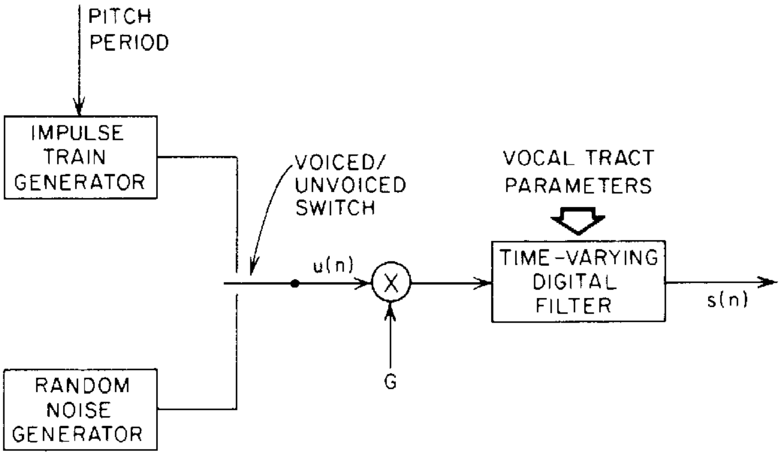


Figure 1.2: The linear source–filter speech production model (from Rabiner and Schafer 1978).

short–time speech analysis is linear prediction (LP). Each speech sample is expressed as a linear combination of the previous samples

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p\, s(n-p) + \text{error} \tag{1.9}$$

where a_i, $i = 1, 2, ..., p$ are the linear prediction coefficients (LPCs) chosen to minimize the prediction error

$$E = \sum_{n=0}^{M} [s(n) - \sum_{i=1}^{p} a_i s(n-p)]^2. \tag{1.10}$$

Here, M is the length in samples of the short time analysis frame and p is the order of the linear predictor (typically chosen to be $2N$, i.e., twice the number of formants/poles). Clearly Eq. (1.9) is equivalent to a filtering procedure; the frequency response of the corresponding all-pole filter is as in Eq. (1.7). The advantages of linear prediction is its simplicity, its mathematical tractability and the existence of fast algorithms for estimating the LPCs. Some disadvantages are the fact that spectral valleys are not modeled and that nonlinear speech production phenomena are ignored. Also, linear prediction is unable to accurately estimate certain speech parameters, e.g., formant bandwidths. Finally, the accuracy of the estimated quantities depends on the choice of the order of the linear predictor (number of poles).

Source–filter modeling and linear prediction are today the dominant approaches in most speech processing applications.[1] Although linear models have been used successfully for many speech applications, they often produces inaccurate or erroneous results.[2] The assumptions that the flow is laminar and filling up the whole vocal tract, that the vocal cavities are passive linear resonators, and that the glottis and vocal tract are uncoupled is only a first order approximation to the complex phenomenon of speech production. Experimental evidence showing that the assumptions of the linear source–filter model are often unjustified is provided in the next section.

1.3 Experimental Evidence in Support of a Nonlinear Model

In [101, 102, 103], Teager has provided strong experimental evidence in support of a nonlinear model. We summarize below the main experimental evidence from aeroacoustics and fluid dynamics that are in disagreement with the linear speech production theory

1. The air flowing in the vocal tract forms jets that often become separated from the surface of the oral cavity and unstable. Teager measured the air velocity and pressure along and across the vocal tract during the phonation of vowels (steady state) and found that the air flow velocity profile in a cross section of the tract is not uniform. The air jet flowing through the vocal tract during the phonation of vowels was found to be unstable and to oscillate between the walls of the mouth.

2. Vortices can easily build up at the lower part of the pharynx [69, 108] and in the front part of the vocal tract [102]. The vortices modulate the air flow, amplifying certain flow frequencies and attenuating others.

3. Numerical simulations of the Navier–Stokes equation for a two dimensional vocal tract [104] and the glottal source [27] support the above presented experimental results. Unfortunately, full scale numerical simulations for a three dimensional vocal tract

[1] Other popular models include the sinusoidal model, multiband filtering analysis–synthesis models and statistical models (e.g., hidden Markov models).

[2] During the past two decades many schemes have been proposed to compensate for the inadequacies of linear models.

have only recently been undertaken due to the very high computational complexity of such simulations.

4. Finally, as the vocal tract shape changes during phonemic transitions, flow instabilities can arise [106].

Teager supplemented his aeroacoustic measurements with a body of work that attempted to explain the above presented experimental results. Unfortunately, in most cases, a rigorous description of the time evolution of the quantities in the "full" Navier–Stokes equation is an impossible task, especially without the help of numerical simulations. Despite these difficulties, Teager provided in his work a rich spectrum of ideas, sometimes labeled as "speech modulation" ideas, based on the dynamical theory of nonlinear oscillators and his intuition on fluid dynamics.

The paper by Kaiser [39] contains an excellent review of references from aeroacoustics, signal processing and fluid dynamics that point out the shortcomings of a linear theory of speech production. In addition, Kaiser also discussed the main "paradoxes" of the widely accepted linear theory. Next, we present evidence in support of a nonlinear model, from a signal processing point of view. Note that few researchers have taken the arduous path of attempting to characterize (or even discuss) nonlinear phenomena during speech production.

In [30], Holmes showed that for real speech the formant amplitude envelope is often quite different from the exponentially decaying amplitude $A\sigma^n$ of Eq. (1.4). Secondary excitations during the open glottal phase were identified as a possible reason for the formant amplitude modulations. It was also noted that the irregularities observed inside a pitch period are not modeled by the linear source–filter model.[3]

In [2], it was shown that source–vocal tract interaction gives rise to a frequency modulation (FM) component in the resonant frequencies, i.e., the formant frequency is different in the open and closed phase of the glottis. The frequency modulation is caused by variations of the damping parameter of the oscillation.[4] Instantaneous changes of the damping

[3] In [4], a multipulse excitation–linear all-pole filter model was proposed, which attempts to model secondary excitations after glottal closure.

[4] The resonant frequency is $\sqrt{\omega^2 - \Gamma^2}$ where ω is the natural frequency of the (undamped) oscillator and Γ is the damping coefficient in the governing differential equation $\ddot{x}(t) + 2\Gamma\dot{x}(t) + \omega^2 x(t) = 0$.

parameter result in changes in the rate of decay of the amplitude envelope, i.e., amplitude modulation (AM). In addition to introducing amplitude and frequency modulation, the source–tract interaction can affect the formant energy levels ("skewing"). An overview of the known effects of the source–tract interaction can be found in [11, 2].

For a fundamental frequency close to the upper or lower limits of the voicing range (e.g., below 80 Hz or over 300 Hz) irregularities appear in the vibration patterns of the vocal cords [28]. In [32], it was shown that a two mass/oscillators model for the vocal cords can produce self–sustained oscillations. Other nonlinear phenomena such as entrainment [72] may also occur for low fundamental frequencies [28]. In general, the excitation can present rapidly time–varying behavior under extreme voicing conditions.

Another piece of evidence in support of a nonlinear speech model is presented in [58, 103]. The energy operator (see Section 2.1) is applied to a speech resonance obtained from bandpass filtering of the speech signal around a formant frequency. According to Eqs. (1.4) and (2.6) the output of the energy operator should be a decaying exponential (at each pitch period), yet, it is shown that more than one "exponentially decaying" pulses exist per pitch period. These "energy pulses" are a strong indication that the vocal cavities often don't behave as linear resonators.

Finally, in [31] departures from the plane wave assumption in the 3–5 kHz frequency range were discussed. Formants in this range were reported to be unstable and with rapidly time–varying characteristics. Amplitude and frequency modulation patterns for a formant in this frequency range can be seen in Fig. 3.7.

An example of the amplitude envelope and instantaneous frequency modulation patterns of a real and of the corresponding synthetic speech resonance are shown in Fig. 1.3, for the third formant $F_3 = 2280$ Hz of the phoneme /eh/ from "thev$\underline{\text{e}}$nin". The speech resonance signal in (b) is obtained by bandpass filtering of the speech signal around the formant frequency (3 dB filter bandwidth is approximately 400 Hz). The synthetic resonance in (a) is synthesized from the simplified linear model of Fig. 1.2 using the parameters estimated from the real speech resonance. The amplitude envelope (c), (d) and instantaneous frequency signals (e), (f) are obtained by demodulating the synthetic and real speech resonances respectively. Note that the rich modulation structure in the real speech resonance is not

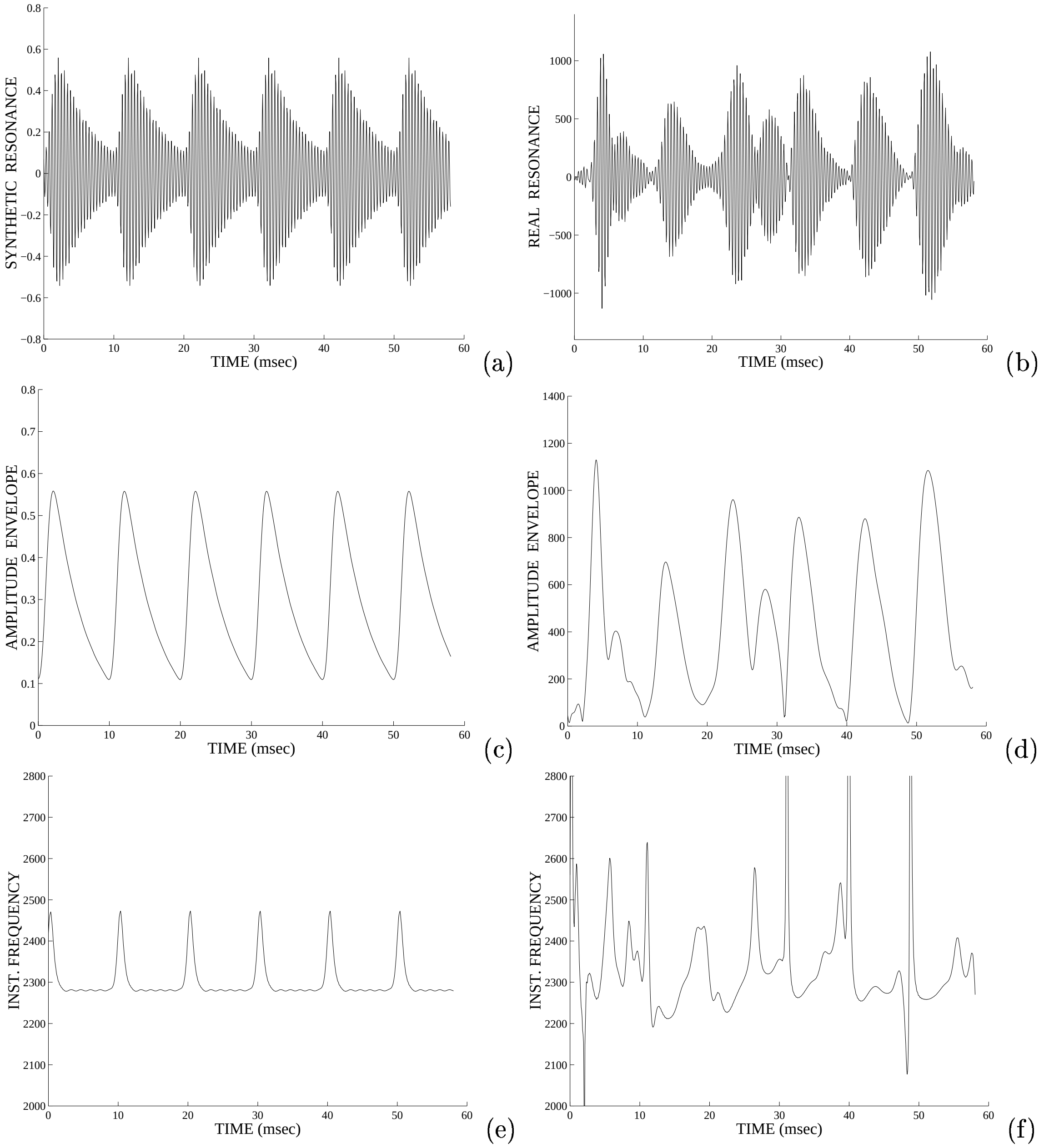


Figure 1.3: Real (b) and resynthesized using the linear source–filter model (a) speech resonance for $F_3 = 2280$ Hz of phoneme /eh/; corresponding amplitude envelope (d), (c) and instantaneous frequency (f), (e) signals (sampling frequency at 16 kHz).

modeled by the linear model (see also Chapter 3 for more examples).

The above presented examples of nonlinear and time–varying phenomena during speech production motivates the introduction of a nonlinear model that attempts to describe the amplitude and frequency modulation structure in speech resonances.

1.4 An AM–FM Speech Modulation Model

Motivated by the above presented experimental results, Maragos, Quatieri and Kaiser [57, 58, 53, 54] proposed a nonlinear model for speech that models each resonance with a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure

$$r(t) = a(t)\cos(\underbrace{2\pi(f_c t + f_m \int_0^t q(\tau)d\tau) + \phi(0)}_{\phi(t)}) \tag{1.11}$$

where $f_c \overset{\triangle}{=} F$ is the formant center frequency, $a(t)$ is the time–varying amplitude signal and $q(t) \in [-1, 1]$ is the frequency modulating/information signal. The instantaneous frequency $f(t)$ is defined as the normalized derivative of the phase $\phi(t)$

$$f(t) = \frac{1}{2\pi}\dot{\phi}(t) = f_c + f_m q(t) \tag{1.12}$$

Finally, f_m is the maximum deviation of the instantaneous frequency from the formant frequency f_c $(0 < f_m < f_c)$.[5]

Using the above formulation one may represent the speech signal $s(t)$ as a sum of N vocal tract resonance signals of Eq. (1.11)

$$s(t) = \sum_{k=1}^{N} r_k(t) \tag{1.13}$$

where N is the number of formants.

The AM–FM modulation model can describe nonlinear and time–varying phenomena during speech production (indirectly) by measuring the modulations present at each speech

[5] Eq. (1.11) does not define $a(t)$ and $f(t)$ uniquely. Certain constraints imposed on the signals $a(t)$ and $f(t)$ (e.g., smoothness) make these quantities physically meaningful and limit the choices of $(a(t), f(t))$ pairs. In the following chapters we will propose algorithms that estimate the amplitude envelope and instantaneous frequency signals from the speech resonance signal $r(t)$. The estimation equations define the $a(t)$ and $f(t)$ signals.

resonance. Most speech modeling efforts are based on the source–filter assumption and attempt to account for speech nonlinearities by using an elaborate source (excitation) model. There, the complicated excitation waveform compensates for the deficiencies of the speech model.[6] As a result, the excitation signal is hard to interpret from a physical point of view and difficult to manipulate or modify, e.g., in a speech synthesis system. Instead of modeling the resonance signal as the output of a source–filter system, the AM–FM model decomposes each formant into amplitude envelope and instantaneous frequency signals. By retaining the excitation–vocal tract coupling, the hard decomposition problem is avoided, at least temporarily, allowing more freedom to observe and quantify the effects of nonlinear speech production phenomena.[7] In addition, the modulation signals have the attractive physical interpretation of being the amount of amplitude and frequency modulation in the speech resonance. Finally, the perceptual importance of amplitude and frequency modulations in speech resonances can be quantified using the AM–FM analysis–synthesis system introduced in Chapter 6.

FM models have been very successful in the past in music analysis and synthesis applications [12]. In [13], Chowning proposed a similar FM model for synthesis of the singing voice, where a single parameter controls the amount of modulation in each speech formant. The AM–FM and the FM models are similar in the sense that they both attempt to describe time–varying and nonlinear phenomena during speech and music production. Yet, the AM–FM model is more general and can represent more complex sounds.

Speech analysis and synthesis using an AM–FM formant model is the open research problem that we attack in this dissertation. Our goal here is to show that the AM–FM model can be successfully applied to speech applications such as formant tracking, pitch estimation, and speech coding. Further, we will provide a parsimonious model for the amplitude envelope and instantaneous frequency formant signals, in the context of a vocoder application. These modeling efforts may have an impact on other areas of speech processing such as speech synthesis and speech recognition/speaker identification.

[6] The advantage of this approach is that the speech analysis–synthesis problem is easier to formulate and solve since a single signal has to be modeled.

[7] A very successful speech model that also does not deconvolve the excitation and vocal tract contributions to the speech signal is the articulatory model [100, 32].

1.5 Organization of the Thesis

The analysis tools of the AM–FM modulation model are introduced in Chapter 2. The energy separation algorithm (ESA) is used to demodulate a speech resonance to amplitude envelope and instantaneous frequency signals. The ESA is based on the energy operator, a differential operator that tracks the energy of the source producing an oscillation. Next, a simple modification to the ESA (smooth ESA) is proposed for increased accuracy and noise robustness. Finally, generalizations of the energy operators to higher orders are discussed and alternate energy demodulation schemes are mentioned.

In Chapter 3, the ESA is compared with the Hilbert transform for AM–FM signal and speech resonance demodulation. The two approaches although fundamentally different (differential vs. integral operators) yield similar results for AM–FM signal and speech resonance demodulation. In addition to the comparisons of the two demodulation approaches, many particularities of the demodulation approach to speech analysis are revealed in this chapter. Specifically the effects of the pitch periodicity and bandpass filtering (needed to isolate a formant) on the formant amplitude envelope and instantaneous frequency signals are discussed qualitatively.

In Chapters 4 and 5, the multiband demodulation formant and pitch tracking algorithms are introduced, based on the AM–FM modulation model, multiband filtering analysis, and ESA demodulation. The speech signal is filtered through a bank of fixed bandpass filters, and each band is demodulated to its amplitude envelope and instantaneous frequency signals. Using the modulation signals, short–time estimates of the formant frequency, formant bandwidth and harmonic frequency are proposed. Finally, the formant and pitch tracking decision algorithms based on the the short–time estimates are presented. Performance, implementation issues, and comparisons with other formant and pitch tracking algorithms conclude each chapter.

In Chapter 6, we present an application of the AM–FM modulation model and the multiband demodulation formant tracker to speech coding. The AM–FM modulation vocoder extracts three or four formant bands from the spectrum by filtering the speech signal along the formant tracks, using bandpass filters with time–varying center frequencies. The formant bands are demodulated to amplitude envelope and instantaneous frequency signals.

These information signals are modeled, coded and quantized. The perceptual importance of the amplitude and frequency modulations in speech resonances is discussed in detail. Finally, relation to other vocoders, implementation, and performance issues are investigated.

Ongoing and future research directions are discussed in Chapter 7. Two very promising applications of the AM–FM modulation model are mentioned: (a) speech synthesis, specifically, style modification of speech using the AM–FM modulation analysis–synthesis system, (b) modulation and fine scale features for speech recognition and speaker identification. In general, the model proposed for the amplitude envelope and instantaneous frequency signals in Chapter 6 provides a general way of quantifying and modifying the amount of amplitude and frequency modulations in speech resonances. This could be potentially useful in many speech applications.

Chapter 2

Energy Operators and AM–FM Demodulation

2.1 Energy Operator

The continuous–time *energy operator* Ψ_c was originally developed by Teager and systematically introduced by Kaiser [37, 38]

$$\Psi_c[x(t)] = \dot{x}(t)^2 - x(t)\ddot{x}(t) \tag{2.1}$$

where dots denote time derivatives. The energy operator tracks the energy[1] of a linear oscillator of mass m and spring of constant k, governed by the differential equation $\ddot{x}(t) + \frac{k}{m}x(t) = 0$, with a general solution $x(t) = A\cos(\omega_0 t + \theta)$, $\omega_0 = \sqrt{k/m}$. The instantaneous energy E_0 of the oscillator is constant and equal to the sum of the kinetic energy of the mass and the potential energy of the spring

$$E_0 = \frac{m}{2}(\dot{x})^2 + \frac{k}{2}x^2 = \frac{m}{2}\,(A\omega_0)^2 \tag{2.2}$$

The energy operator applied on the displacement $x(t)$ gives

$$\Psi_c[A\cos(\omega_0 t + \theta)] = A^2{\omega_0}^2 = \frac{E_0}{m/2} \tag{2.3}$$

[1] Note that the energy required to produce an oscillation is different from the "energy" that is defined in most signal processing textbooks as the short–time average of the signal squared. To differentiate, we refer to the later as the "classic energy."

that is the product of the amplitude and the frequency squared or equivalently the energy of the oscillator per unit mass. The first term of Ψ_c, $\dot{x}(t)^2$ is proportional to the kinetic energy, while the second term $x(t)\ddot{x}(t)$ is proportional to the potential energy of the oscillation. In general, for a damped cosine with damping coefficient Γ

$$\Psi_c[Ae^{\Gamma t}\cos(\omega_0 t+\theta)] = A^2 e^{2\Gamma t}{\omega_0}^2. \tag{2.4}$$

In discrete–time, the energy operator

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{2.5}$$

is derived from Ψ_c by approximating derivatives with forward or backward difference operators [55]. An interesting alternative way of discretizing Ψ_c is by shifting the forward differences so that the products $\dot{x}(t)^2$ and $x(t)\ddot{x}(t)$ in Ψ_c are computed at the same sample n, i.e.,

$$\begin{aligned}\Psi_a[x(n)] &= [x(n)-x(n-1)]\,[x(n+1)-x(n)] - x(n)\,[x(n-1)-2x(n)+x(n+1)] \\ &= x^2(n)-x(n+1)x(n-1) = \Psi_d[x(n)].\end{aligned}$$

For a discrete damped cosine with damping coefficient σ

$$\Psi_d[A\sigma^n\cos(\Omega_0 n+\theta)] = A^2\sigma^{2n}\sin^2(\Omega_0). \tag{2.6}$$

2.1.1 Energy Separation Algorithm

In [58, 55, 53, 54], Maragos, Kaiser, and Quatieri applied the energy operators Ψ_c and Ψ_d to AM–FM signal demodulation. A real–valued AM–FM signal as a modulated cosine is defined as

$$x(t) = a(t)\cos(\underbrace{\omega_c t + \omega_m\int_0^t q(\tau)d\tau + \phi(0)}_{\phi(t)}) \tag{2.7}$$

with a time–varying amplitude signal $a(t)$, and a time–varying instantaneous angular frequency signal

$$\omega(t) \triangleq \frac{d\phi}{dt}(t) = \omega_c + \omega_m q(t), \tag{2.8}$$

where ω_c is the carrier frequency, $q(t) \in [-1,1]$ is the frequency modulating signal, $\omega_m \in [0,\omega_c]$ is the maximum frequency deviation, and $\phi(0)$ is an arbitrary phase offset. A typical demodulation problem is, given $x(t)$, to estimate the *amplitude envelope* $|a(t)|$ and *instantaneous frequency* $\omega(t)$ signals. Clearly there is not a unique choice of $|a(t)|$ and $\omega(t)$ that satisfies Eq. (2.7). Assuming that $|a(t)|$ and $\omega(t)$ are arbitrary bandlimited signals with bandwidth much smaller that the carrier frequency ω_c, strict error bounds can be provided for the amplitude envelope and instantaneous frequency estimation errors, as discussed next.

When Ψ_c is applied to the AM–FM signal $x(t)$ it can approximately estimate the squared product of the amplitude and frequency signals; i.e.,

$$\Psi_c[x(t)] \approx [a(t)\omega(t)]^2 \tag{2.9}$$

assuming that the signals $a(t)$ and $\omega(t)$ do not vary too fast or too greatly in time compared to the carrier frequency ω_c [55]. To separate the above "energy product" the following *Energy Separation Algorithm (ESA)* was developed in [53, 54] by Maragos, Kaiser and Quatieri

$$\omega_E(t) = \sqrt{\frac{\Psi_c[\dot{x}(t)]}{\Psi_c[x(t)]}} \quad \approx \quad \omega(t) \tag{2.10}$$

$$|a_E(t)| = \frac{\Psi_c[x(t)]}{\sqrt{\Psi_c[\dot{x}(t)]}} \quad \approx \quad |a(t)|. \tag{2.11}$$

At each time instant the ESA estimates the instantaneous frequency and the amplitude envelope of x by using only the two instantaneous output values of the energy operator applied to the signal x and its derivative $\dot{x}$. Upper bounds for the approximation errors in (2.9) can be expressed in terms of the ratios of the bandwidth of $a(t)$, $\omega(t)$ and the carrier frequency ω_c [58, 53, 55]. If $x(t) = A\cos(\omega_c t)$ is a cosine with no AM or FM, the ESA yields exact estimates of the constant amplitude and frequency, i.e., $\omega_E = \omega_c$ and $|a_E| = |A|$.

Using Ψ_d similar demodulation methods can be applied to discrete–time AM–FM signals

$$x(n) = a(n)\cos\underbrace{(\Omega_c n + \Omega_m \int_0^n q(k)dk + \phi(0))}_{\phi(n)} \tag{2.12}$$

to estimate their amplitude envelope $|a(n)|$ and angular instantaneous frequency

$$\Omega(n) = \frac{d\phi}{dn}(n) = \Omega_c + \Omega_m q(n)$$

where $0 \le \Omega_m \le \Omega_c$ and $|q(n)| \le 1$. Note that the continuous–time frequencies ω_c, ω_m, and $\omega(t)$ have been replaced by their discrete–time counterparts Ω_c, Ω_m, and $\Omega(n)$, which are assumed to be in $[0, \pi]$. If $x(n)$ has resulted from sampling a continuous–time signal, then

$$\Omega_c = \omega_c T \ , \ \Omega_m = \omega_m T \ , \ \Omega = \omega T$$

where T is the sampling period. It has been shown in [58, 55] that

$$\Psi_d[x(n)] \approx a^2(n)\sin^2[\Omega(n)]. \tag{2.13}$$

By applying Ψ_d to both $x(n)$ and its backward difference

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

a discrete–time ESA has been developed in [53, 54]:

$$\Omega_E(n) \ = \ \arccos\left(1 - \frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4\Psi_d[x(n)]}\right) \ \approx \ \Omega(n) \tag{2.14}$$

$$|a_E(n)| = \sqrt{\frac{\Psi_d[x(n)]}{1 - \left(1 - \frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4\Psi_d[x(n)]}\right)^2}} \ \approx \ |a(n)| \tag{2.15}$$

The frequency estimation part assumes that $0 < \Omega(n) < \pi$. Thus, the discrete ESA algorithm can estimate instantaneous frequencies up to half of the sampling frequency. The approximations in (2.13) and (2.14),(2.15) are valid under similar assumptions as in the continuous–time case. Examples of AM–FM signal and speech resonance demodulation using the ESA can be found in Chapter 3.

2.1.2 Smoothed Energy Separation Algorithm

The precise result from applying the energy operator to an AM–FM signal is (from [58])

$$\Psi_c[a(t)\cos(\phi(t))] = \overbrace{(a\dot{\phi})^2}^{D} + \overbrace{\frac{\Psi_c(a)}{2}}^{E_L} + \overbrace{\frac{a^2\ddot{\phi}}{2}\sin(2\phi) + \frac{\Psi_c(a)}{2}\cos(2\phi)}^{E_H}. \tag{2.16}$$

The desired term is $D = (a\dot{\phi})^2$, whereas E_L and E_H are the error terms in the approximation (2.9). Note that the energy operator approximation incurs a low–frequency error component E_L and a high–frequency component E_H concentrated around $2\,\omega_c$, that is twice the carrier frequency of the AM–FM signal. If the amplitude envelope and instantaneous frequency signals are bandlimited with a highest frequency that is much smaller than ω_c then the

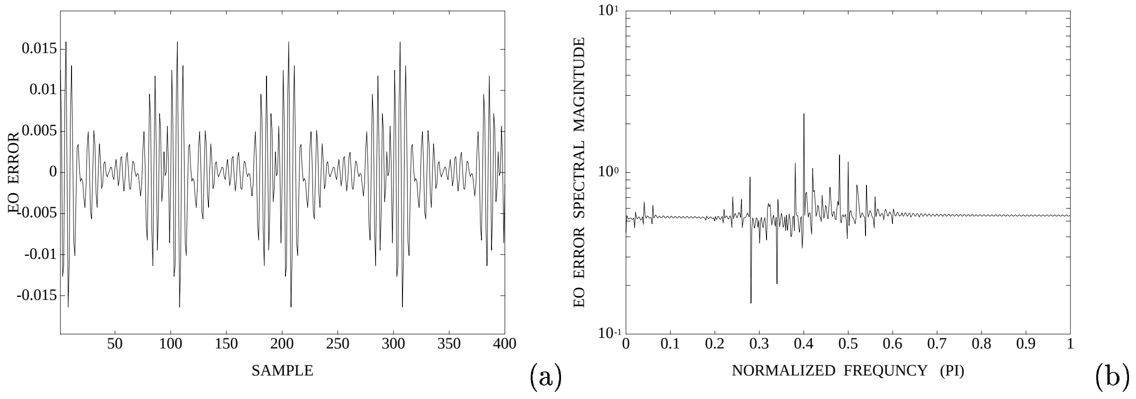


Figure 2.1: (a) The energy operator approximation error for the AM–FM signal $x(n) = (1+0.5\cos(\pi n/50))\cos[\pi n/5+\sin(\pi n/25)+\phi]$. (b) The magnitude of the Fourier transform of the approximation error.

bandwidth of the desired term D is also much smaller than ω_c. In this case, the high–frequency error component E_H is well separated from the desired term D in the frequency domain. Thus, by filtering the energy operator output through an appropriate lowpass filter, one can eliminate the high–frequency error component E_L without affecting the low–frequency desired term D.

Similarly, in discrete–time, when the energy operator Ψ_d is applied to an AM–FM signal, we get a high–frequency error component concentrated around $2\,\Omega_c$ as shown in Fig. 2.1. The approximation error signal of Eq. (2.13) and its Fourier spectrum are displayed in Fig. 2.1 (a) and (b) respectively, for the AM–FM/cosine signal $x(n) = (1 + 0.5\cos(\pi n/50))\cos[\pi n/5 + \sin(\pi n/25) + \phi]$. Clearly, the error has a high frequency component around $2\,\Omega_c = 0.4\pi$ that can be eliminated through low–pass filtering.

The choice of an appropriate lowpass filter is not straightforward. Clearly, an "expensive" filter with a long impulse response can decrease considerably the approximation error. We must keep in mind though, that one of the major advantages of the energy operator is its instantaneous nature, which results in excellent time resolution. This property is valuable for applications where the information signals may have abrupt transitions (e.g., pitch or phonemic transitions in speech signals). To conserve the instantaneous nature and the simplicity of Ψ_d we choose a filter with a short impulse response. A 7–point linear binomial

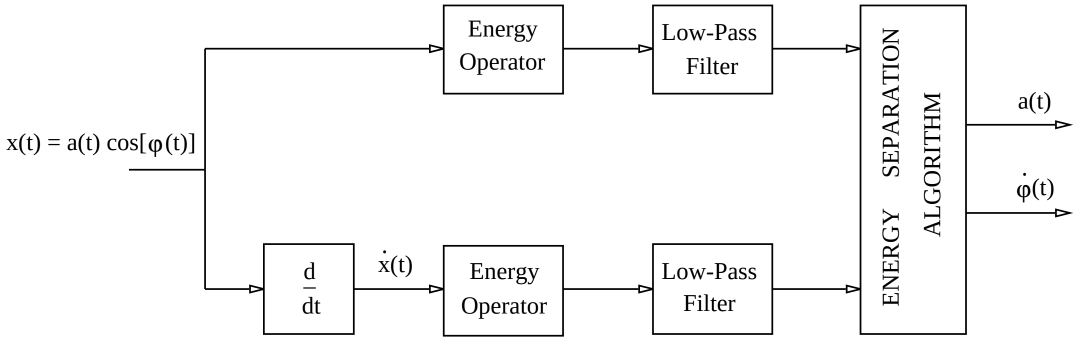


Figure 2.2: The block diagram of the Smoothed Energy Separation Algorithm.

smoothing filter with impulse response (1, 6, 15, 20, 15, 6, 1) is used for sampling frequencies in the 16–20 kHz range.[2] This filter is equivalent to the 3–point filter (1, 2, 1) applied to the energy operator output three times or to the 2–point moving average filter (1, 1) applied six times. With this simple and computationally inexpensive smoothing, the energy operator approximation error decreases typically by 50%. The envelope and frequency estimation errors are also reduced when the smoothed energy signals are used in the separation algorithm. Henceforth, we refer to the envelope and frequency separation algorithm using the binomially smoothed energy signals, as the *Smoothed Energy Separation Algorithm* (SESA), summarized in Fig. 2.2.

In [55], it is shown that the discrete–time operator Ψ_d is obtained from the continuous–time operator Ψ_c by approximating $\dot{x}(t)$ by $x(n) - x(n-1)$. Further, Ψ_d followed by a 3–point binomial filter (1, 2, 1) is equivalent to using the 3–sample symmetric difference $[x(n+1) - x(n-1)]/2$ when discretizing $\dot{x}(t)$. Approximations of $\dot{x}(t)$ that involve more samples offer an alternative way of improving the performance of the energy operator, with results similar to the binomial smoother.

Finally, smoothing can be applied on the estimated information signals (post–smoothing) instead of the energy signals (pre–smoothing), because the envelope and frequency estimation error signals have a high frequency component around $2\,\omega_c$, very much like the energy operator error does. Both approaches (post and pre–smoothing) yield similar error reductions. There are applications where post–smoothing is advantageous, e.g., when median

[2] Smoothing the output of the energy operator via lowpass filtering has also been implemented in [87], when using the energy operator as a detector of transient signal signatures in AM–FM background noise.

filtering is also performed.

2.2 Higher–Order Energy Operators

In [56], the energy operator is formalized in a generalized higher–order differential energy operator class. Instantaneous differences in the relative rate of change between two signals x, y can be measured via their Lie bracket $[x,y] \overset{\triangle}{=} \dot{x}y - x\dot{y}$, because $[x,y]/xy = (\dot{x}/x) - (\dot{y}/y)$. If $y = \dot{x}$, then $[x,y]$ becomes the energy operator $\Psi_c = [x, \dot{x}]$. In the general case, if x and y represent displacements in some generalized motions, the quantity $[x, \dot{y}] = \dot{x}\dot{y} - x\ddot{y}$ has dimensions of energy (per unit mass), and hence we may view it as a *"cross energy"* between x and y.

The k^{th}–order *differential energy operator* (DEO) is defined in [56] as

$$\Upsilon_k(x) \overset{\triangle}{=} [x, x^{(k-1)}] = \dot{x}x^{(k-1)} - xx^{(k)}, \qquad k = 0, \pm 1, \pm 2, ... \tag{2.17}$$

and yields the cross energy between a signal $x(t)$ and its $(k-1)^{\text{th}}$ derivative (or integral). Clearly, $\Upsilon_2 \equiv \Psi_c$. The third–order DEO $\Upsilon_3 \overset{\triangle}{=} \dot{x}\ddot{x} - xx^{(3)}$ is an *energy velocity* operator, whereas the fourth–order DEO $\Upsilon_4 \overset{\triangle}{=} \dot{x}x^{(3)} - xx^{(4)}$ has dimensions of *energy acceleration.* For a sine wave

$$\Upsilon_k[A\cos(\omega_0 t + \theta)] = \begin{cases} 0, & k = \pm 1, \pm 3, \pm 5, ... \\ (-1)^{1+\frac{k}{2}} A^2 \omega_0^k, & k = 0, \pm 2, \pm 4, ... \end{cases} \tag{2.18}$$

Similar "energy equations" hold approximately for AM–FM signals provided that the amplitude envelope $a(t)$ and the instantaneous frequency $\omega(t)$ signals do not vary too fast or too much with respect to the carrier frequency. Further, because $a^2(t)\omega^k(t)$ are lowpass signals, the above instantaneous energy measures can be used for robust estimation of instantaneous amplitude and frequency in modulated sinusoids.

In discrete time, the k^{th}–order energy operator[3] is defined as

$$\Upsilon_k(x[n]) = x[n]\, x[n+k-2] - x[n-1]\, x[n+k-1], \qquad k = 0, 1, 2, 3, ... \tag{2.19}$$

For $k = 2$ we obtain the standard discrete energy operator $\Upsilon_2 \equiv \Psi_d$. For $k = 3$ we obtain an *asymmetric discrete energy velocity operator* $\Upsilon_3(x_n) \overset{\triangle}{=} x_n x_{n+1} - x_{n-1}x_{n+2}$, whereas $k = 4$

[3] For notational simplicity Υ_k is used for both continuous- and discrete–time higher–order energy operators.

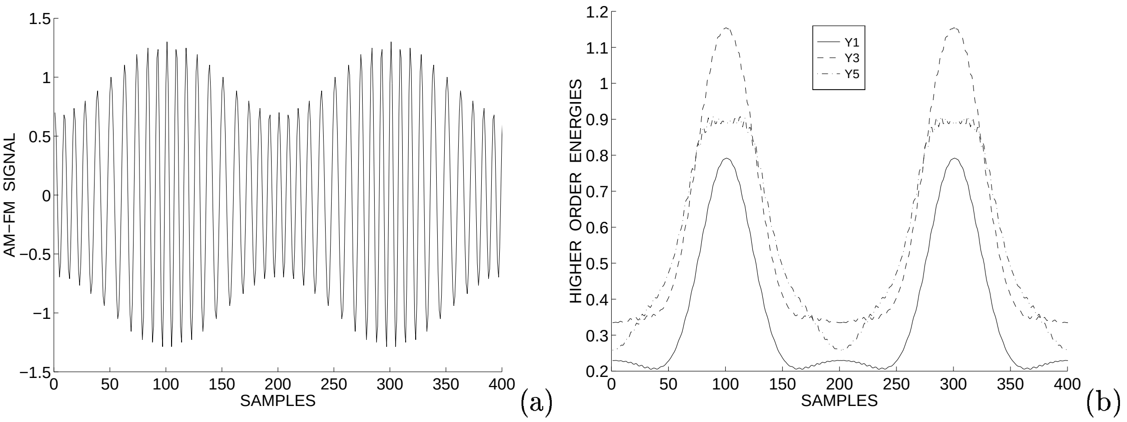


Figure 2.3: (a) The AM–FM signal $x(n) = (1 + 0.3\cos(\pi n/50))\cos[\pi n/5 + 2\sin(\pi n/50) + \phi]$ and (b) the first-, third- and fifth–order energy operator outputs: Υ_1 (solid), Υ_3 (dashed) and Υ_5 (dashed–dotted).

yields a discrete energy acceleration operator $\Upsilon_4(x_n) \overset{\triangle}{=} x_n x_{n+2} - x_{n-1}x_{n+3}$. Applying the operators Υ_k to discrete (possibly damped) cosines yields discrete energy equations

$$\Upsilon_k[A\sigma^n \cos(\Omega_0 n + \theta)] = A^2\sigma^{2n+k-2}\sin(\Omega_0)\sin[(k-1)\Omega_0] \tag{2.20}$$

which are useful for parameter estimation in sinusoids. In addition, Eq. (2.20) holds approximately when the cosine has slowly time–varying amplitude and frequency, i.e., for a sampled AM–FM signal. This allows us to find discrete AM–FM demodulation algorithms by combining the above energy equations of various orders. For example, by using Υ_2, Υ_3, and the undamped cosine energy equations $\Upsilon_k[A\cos(\Omega_0 n + \theta)] = A^2\sin(\Omega_0)\sin[(k-1)\Omega_0)]$ for $k = 2, 3$, a discrete algorithm was proposed in [17] for instantaneous frequency tracking, which is closely related to the discrete ESA. The higher–order energy operators have been used successfully in co–channel demodulation and feature extraction for sums of AM–FM signals [95].

In Fig. 2.3 (b), we display the output of the first-, third- and fifth–order energy operators for the AM–FM signal shown in (a). Note that all three energy operators, Ψ, Υ_3, and Υ_5, produce slow–varying output.

2.3 Quadratic Discrete Energy Operators

The good performance of the ESA is mainly due to the fact that the output of Ψ_d is slowly varying for AM–FM signals and constant for sinusoids, i.e., the energy signals $\Psi_d[x(n)]$ and $\Psi_d[x(n) - x(n-1)]$ have small bandwidth. Time–invariance to sinusoidal input is a very useful property for operators when applied to AM–FM demodulation problems. A general two–parameter family of *quadratic energy operators* Q_{km} that yield time–invariant output for a sinusoidal input is

$$Q_{km}(x[n]) \overset{\triangle}{=} x[n]x[n+k] - x[n-m]x[n+k+m], \qquad k = 0, 1, 2, ..., \quad m = 1, 2, ... \tag{2.21}$$

For $k = 0$, $m = 1$, the energy operator Ψ_d is obtained. Q_{km} is the most general subclass of quadratic energy operators of the form $x(n)x(n+k) - x(n+l)x(n+m)$, $k, l, m \in \mathcal{Z}$, that produce time–invariant output for a sinusoidal input. The general energy equations for Q_{km} are:

$$Q_{km}[A\sigma^n \cos(\Omega_0 n + \theta)] = A^2 \sigma^{2n+k} \sin(m\Omega_0) \sin[(m+k)\Omega_0]. \tag{2.22}$$

Similar equations hold for AM–FM signals, provided the modulation signals $a(t)$ and $\omega(t)$ do not vary too fast or too much with respect to the carrier frequency. The quadratic energy operators can be combined to produce AM–FM demodulation algorithms. Note, that each Q_{km} can be generated recursively from operators of lower orders k, m.

The family of Q_{km} operators has also been studied independently by Kaiser. Further, the discrete higher–order energy operator class Υ_k are a subset of the quadratic operators Q_{km} since $Q_{k1} \equiv \Upsilon_{k+2}$; e.g., $Q_{01} \equiv \Psi_d$ and $Q_{11} \equiv \Upsilon_3$. For $k = 0$ the operators Q_{0m} can also be viewed as a special case of the class $\sum_m h_m x[n+m]x[n-m]$ of quadratic detectors proposed in [5].

Of interest is also the most general quadratic sum of the form

$$\sum_{i=-I}^{i=I} \sum_{j=-J}^{j=J} a_{ij} \; x(n+i) \; x(n+j) \tag{2.23}$$

with arbitrary weights a_{ij}. We solve for weights a_{ij} that result in operators that are time–invariant to sinusoidal input. Any operator produced from the general sum of Eq. (2.23) for $I, J \leq 2$ (i.e., windows up to 5 samples long) can be expressed as linear combination of the quadratic energy operators Q_{km}.

Chapter 3

Comparisons of Energy Separation and Hilbert Demodulation

3.1 Introduction

A typical demodulation problem is to estimate the amplitude envelope $|a(t)|$ and instantaneous frequency $\omega(t)$ signals, given the AM–FM signal $x(t)$ defined in Eq. (2.7). As we have already discussed the above–posed demodulation problem is ill–defined, because an infinite number of $(|a(t)|, \omega(t))$ pairs exist that satisfy Eq. (2.7). Additional smoothness constraints imposed on the amplitude envelope and instantaneous frequency signals make the problem mathematically tractable. Error bounds can be provided for the $|a(t)|$ and $\omega(t)$ estimates of the demodulation algorithms as discussed next.

A standard approach to this demodulation problem is to use the Hilbert transform and the related Gabor's analytic signal [22, 97, 78]. An alternative approach to demodulation is the energy separation algorithm (ESA, SESA) introduced in Section 2.1. The ESA uses the energy operator (Eq. (2.1)) to first estimate the energy required to produce the AM–FM signal, and then separate the energy into its amplitude and frequency components.

In this chapter, we compare these two fundamentally different approaches to AM–FM signal demodulation. The Hilbert transform approach (reviewed in Section 3.2.2) mainly involves a linear integral operator, whereas the energy operator approach uses a nonlinear differential operator. In Section 3.3, the demodulation algorithms are applied to arbitrary synthetic signals. The magnitude of estimation errors, the computational complexity, the behavior in the presence of noise, and the time–resolution of the two algorithms is compared.

A promising application area for the methods compared in this chapter is the problem of tracking amplitude and frequency modulations in speech resonances. As discussed in Section 1.4, the AM–FM modulation speech model was introduced by Maragos, Quatieri, and Kaiser [57, 58] in an attempt to describe several nonlinear and time–varying phenomena during speech production. The modulation model represents a single speech resonance (formant) within a pitch period as a damped AM–FM signal. Estimating the amplitude envelope and the instantaneous frequency of the resonance signal involves a demodulation algorithm.

Before applying either demodulation approach to a single speech resonance signal, the resonance needs to be isolated by *bandpass filtering* of the speech signal. The main effect of bandpass filtering is to blur the fine time detail of the amplitude and frequency modulating signals. In Section 3.4.2, following the work of Papoulis on the subject [80], the blurring imposed by the filter on a general AM–FM input signal is analyzed. *Pitch periodicity* of vowel speech signals also poses certain problems to AM–FM modeling and demodulation. Hence, in Section 3.4, we compare experimentally the effects of bandpass filtering and pitch periodicity when the Hilbert transform or the ESA is applied to speech vowel resonance demodulation. To clarify the underlying ideas from the experiments on real speech, several comparisons on *synthetic* speech vowels are presented in Section 3.4. We have found that the conclusions drawn from experiments on synthetic and real speech are similar. Furthermore, since the parameters of the synthetic speech signals are known, the accuracy of the various demodulation approaches can be easily verified. Finally, in Section 3.5, we summarize our conclusions from the comparisons of the two AM–FM demodulation approaches.

3.2 Amplitude/Frequency Separation Algorithms

Consider the real–valued AM–FM signal $x(t) = a(t)\cos[\phi(t)]$ in Eq. (2.7). By "amplitude/frequency separation" or "demodulation" we refer to the estimation of the amplitude envelope $|a(t)|$ and the instantaneous frequency $\omega(t) = \dot{\phi}(t)$, which we call *information sig-*

nals.[1] Assuming that ω_c is known, estimating $\omega(t)$ is equivalent to estimating the frequency modulating signal $q(t)$. Similarly, if $a(t) \geq 0$ for all t, one can write $a(t) = A[1 + \kappa b(t)]$ where $0 \leq \kappa \leq 1$, $|b(t)| \leq 1$, and A is some positive constant; estimating $|a(t)|$ is equivalent to estimating the amplitude modulating signal $b(t)$. In general, amplitude signals $a(t)$ may be nonnegative; however, for the purposes of this thesis we consider it sufficient to estimate the amplitude envelope.

3.2.1 Energy Separation Algorithms

The energy separation algorithm (ESA) was introduced in Section 2.1.1 to demodulate a real AM–FM signal to its amplitude envelope and instantaneous frequency components. The continuous- and discrete–time ESAs are summarized by Eqs. (2.10), (2.11) and Eqs. (2.14), (2.15) respectively. The smoothed energy separation algorithm (SESA) was introduced in Section 2.1.2 and summarized in Fig. 2.2. A 7–point binomial smoothing filter is applied on the energy signals for this SESA implementation. In the comparisons among demodulation algorithms performed in this chapter, the discrete–time ESA and SESA are being used.

3.2.2 Hilbert Transform Demodulation

The *Hilbert transform* of any signal $x(t)$ is

$$\hat{x}(t) = x(t) * \frac{1}{\pi t} \tag{3.1}$$

with Fourier transform

$$\hat{X}(\omega) = -j\,\mathrm{sgn}(\omega)\,X(\omega) \tag{3.2}$$

where $X(\omega)$ is the Fourier transform of $x(t)$. The *analytic signal* of $x(t)$ is defined as

$$z(t) = x(t) + j\hat{x}(t) = |z(t)|\ \exp[j\theta(t)]. \tag{3.3}$$

For an AM–FM signal $x(t)$ of the form

$$x(t) = a(t)\cos[\phi(t)] \tag{3.4}$$

[1] As noted above certain smoothness constraints are imposed on the information signals (depending on the demodulation algorithm) to make the estimation problem mathematically tractable.

the modulus $|z(t)|$ and phase derivative $\dot{\theta}(t)$ of the analytic signal can serve as (generally approximate) estimates for the amplitude envelope and instantaneous frequency of $x(t)$. Thus the *Hilbert transform demodulation algorithm* (HTD) is given by the following two equations:

$$\begin{aligned} |a_H(t)| &= \sqrt{x^2(t) + \hat{x}^2(t)} \;\approx\; |a(t)| && (3.5) \\ \omega_H(t) &= \frac{d}{dt}(\arctan\left[\frac{\hat{x}(t)}{x(t)}\right]) \;\approx\; \omega(t). && (3.6) \end{aligned}$$

Let the *quadrature signal* of $x(t)$ be defined as

$$x_q(t) = a(t)\sin[\phi(t)]. \tag{3.7}$$

Clearly, if the Hilbert transform of $x(t)$ is equal to its quadrature signal, then the HTD estimates $r(t)$ and $\dot{\theta}(t)$ are equal to the actual information signals $|a(t)|$ and $\omega(t)$. In general, though, $\hat{x}(t)$ and $x_q(t)$ are not equal, thus an envelope $e_a(t)$ and frequency $e_\omega(t)$ estimation error is present. These estimation errors are closely related to the *quadrature error* signal of the Hilbert transform defined as

$$e(t) = x_q(t) - \hat{x}(t) = a(t)\sin[\phi(t)] - \hat{x}(t). \tag{3.8}$$

Consider the complex–valued signal

$$w(t) = x(t) + jx_q(t) = a(t)\exp[j\phi(t)] \tag{3.9}$$

with Fourier transform $W(\omega)$. Then

$$X(\omega) = \frac{1}{2}[W(\omega) + W^*(-\omega)]. \tag{3.10}$$

Nuttall [75, 76] has shown that the total "classic energy" in the quadrature error signal is

$$E = \int_{-\infty}^{+\infty} |e(t)|^2 dt = \frac{1}{\pi}\int_{-\infty}^{0} |W(\omega)|^2 d\omega. \tag{3.11}$$

Therefore, if $W(\omega)$ is zero for negative frequencies the quadrature error $e(t)$ is also zero. When $W(\omega)$ extends to negative frequencies, the quadrature error $e(t)$ is non–zero and the magnitude of the error increases as the negative side of $W(\omega)$ grows. In the special case of a cosine $x(t) = A\cos(\omega_c t)$, the HTD provides exact estimates of the amplitude and frequency, because $\hat{x}(t) = A\sin(\omega_c t)$ and hence $e(t) = 0$ for all t.

The quadrature error $e(t)$ can provide bounds for the estimation errors of the information signals. The envelope estimation error $e_a(t)$ may be expressed as:

$$\begin{aligned}
e_a(t) &= |a(t)| - |a_H(t)| = |a(t)| - \sqrt{x^2(t) + \hat{x}^2(t)} \\
&= |a(t)| - \sqrt{x^2(t) + [x_q(t) - e(t)]^2} \\
&= |a(t)| \left(1 - \sqrt{1 - 2\frac{e(t)}{a(t)} \sin[\phi(t)] + \frac{e^2(t)}{a^2(t)}}\right) \\
&\approx e(t)\, \mathrm{sgn}[a(t)] \sin[\phi(t)] - \frac{e^2(t)}{2|a(t)|}, \quad \text{if} \quad |e(t)| \ll |a(t)|,\ a(t) \neq 0. \qquad (3.12)
\end{aligned}$$

The phase estimation error is

$$\begin{aligned}
e_\phi(t) &= \phi(t) - \arctan\left[\frac{\hat{x}(t)}{x(t)}\right] \\
&= \phi(t) - \arctan\,[\,\tan[\phi(t)] - \frac{e(t)}{a(t)\cos[\phi(t)]}\,]
\end{aligned}$$

Hence:

$$\begin{aligned}
\tan[\phi(t) - e_\phi(t)] = \tan[\phi(t)] - \frac{e(t)}{a(t)\cos[\phi(t)]} \quad &\Longrightarrow \\
\tan[e_\phi(t)] = \frac{e(t)\cos[\phi(t)]}{a(t) - e(t)\sin[\phi(t)]} \quad &\Longrightarrow \\
e_\phi(t) \approx \frac{e(t)\cos[\phi(t)]}{a(t)}\,, \qquad \text{if} \quad |e(t)| \ll |a(t)|,\ \ a(t) \neq 0. \qquad &(3.13)
\end{aligned}$$

Finally, the instantaneous frequency estimation error e_ω is simply the derivative of the phase error e_ϕ. In brief, for $|e(t)| \ll |a(t)|$ and $a(t) \neq 0$

$$|e_a(t)| \leq |e(t)| \qquad (3.14)$$

$$|e_\phi(t)| \leq \left|\frac{e(t)}{a(t)}\right| \qquad (3.15)$$

$$|e_\omega(t)| = |\omega(t) - \omega_H(t)| = |\dot{e}_\phi(t)| \approx \left|\frac{d}{dt}\left[\frac{e(t)\cos[\phi(t)]}{a(t)}\right]\right|. \qquad (3.16)$$

For the discrete–time signal $x(n)$ its Hilbert transform $\hat{x}(n) = x(n) * h(n)$ is defined [78] in the time domain as the convolution of $x(n)$ with the infinite impulse response

$$h(n) = \begin{cases} \frac{2}{\pi}\frac{\sin^2(\pi n/2)}{n} & n \neq 0 \\ 0 & n = 0 \end{cases} \qquad (3.17)$$

In practice, one can implement the Hilbert transform by using a finite impulse response (FIR) approximation to the infinite impulse response (IIR) $h(n)$. Such FIR filter designs can be obtained either via the window method (e.g. Kaiser windows) or the equiripple method [78]. An alternative way of approximating the discrete–time analytic signal $z(n) = x(n) + j\hat{x}(n)$ is by using the fast Fourier transform (FFT) to implement a 90^o phase splitter [78].

If $x(n) = a(n)\cos[\phi(n)]$ and $x_q(n) = a(n)\sin[\phi(n)]$, the total "classic energy" of the quadrature error becomes in discrete time

$$E = \sum_{n=-\infty}^{\infty} |x_q(n) - \hat{x}(n)|^2 = \frac{1}{\pi}\int_{-\pi}^{0} |W(\Omega)|^2 d\Omega \tag{3.18}$$

where $W(\Omega)$ is the discrete Fourier transform of the signal $w(n) = x(n) + jx_q(n)$. The envelope and instantaneous frequency estimation equations (3.5), (3.6) and error bound equations (3.14), (3.15), (3.16) also hold in discrete time. However, in addition to the quadrature error that depends on the signal $x(n)$, any (FIR or FFT) discrete Hilbert transform implementation also incurs an additional error, by being an approximation of the exact IIR Hilbert transformer.

In this chapter, we will use two FIR Hilbert transformers designed via the window method: (i) a filter with a short 19–sample impulse response, cutoff frequency at 500 Hz, and 10% maximum ripple in the passband; (ii) a filter with a long 139–sample impulse response, cutoff frequency at 200 Hz, and 1% ripple. The two Hilbert transform demodulation algorithms corresponding to the above implementations will be referred to as "short HTD" and "long HTD", respectively. Both implementations assume a sampling frequency of 20 kHz.

3.3 Comparisons on Synthetic Signals

In this section, the three amplitude/frequency separation algorithms (ESA, SESA, and HTD) are compared using discrete–time AM–FM/cosine signals $x(n)$ of the form:

$$x(n) = (1 + \kappa\cos(W_a\, n))\cos[\Omega_c n + (\Omega_m/W_f)\sin(W_f\, n)] \tag{3.19}$$

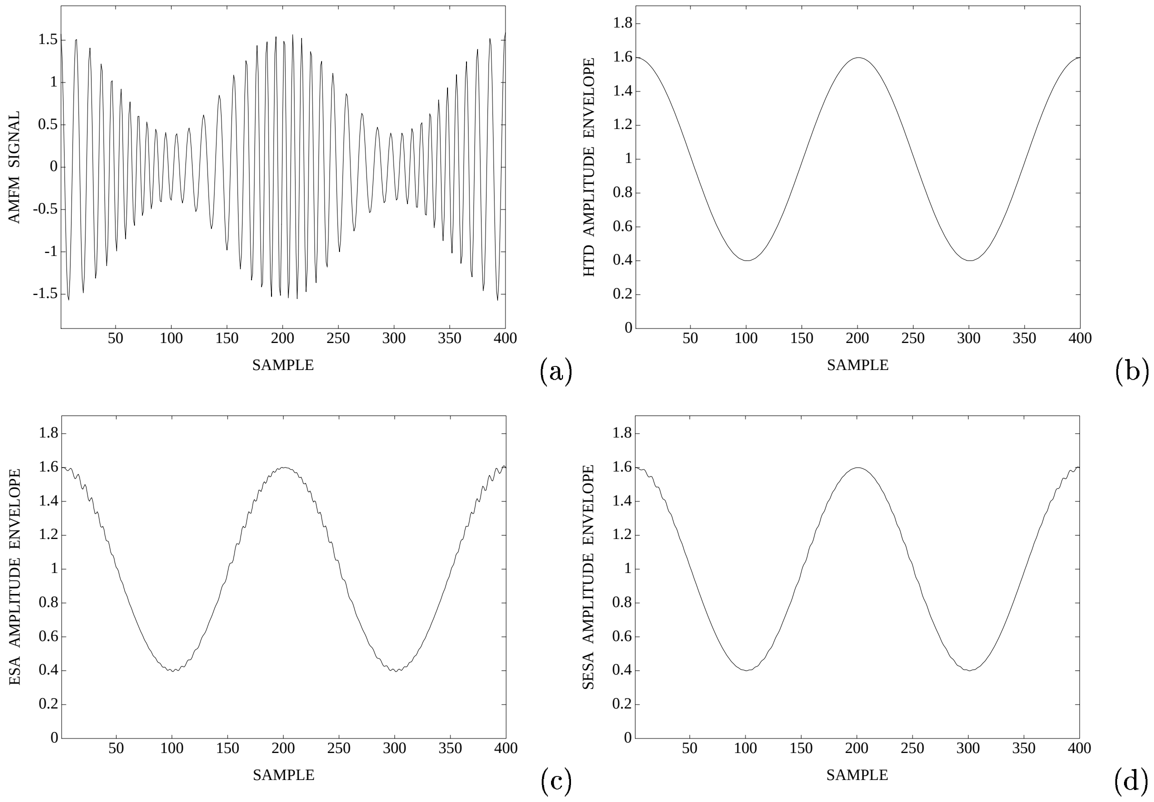


Figure 3.1: (a) AM–FM signal $x(n) = (1 + 0.6\cos[\pi n/100])\cos[\pi n/5 + 4\sin(3\pi n/200)]$. Estimated amplitude envelope using the: (b) Hilbert Transform Demodulation Algorithm (HTD), (c) Energy Separation Algorithm (ESA), and (d) Smoothed Energy Separation Algorithm (SESA).

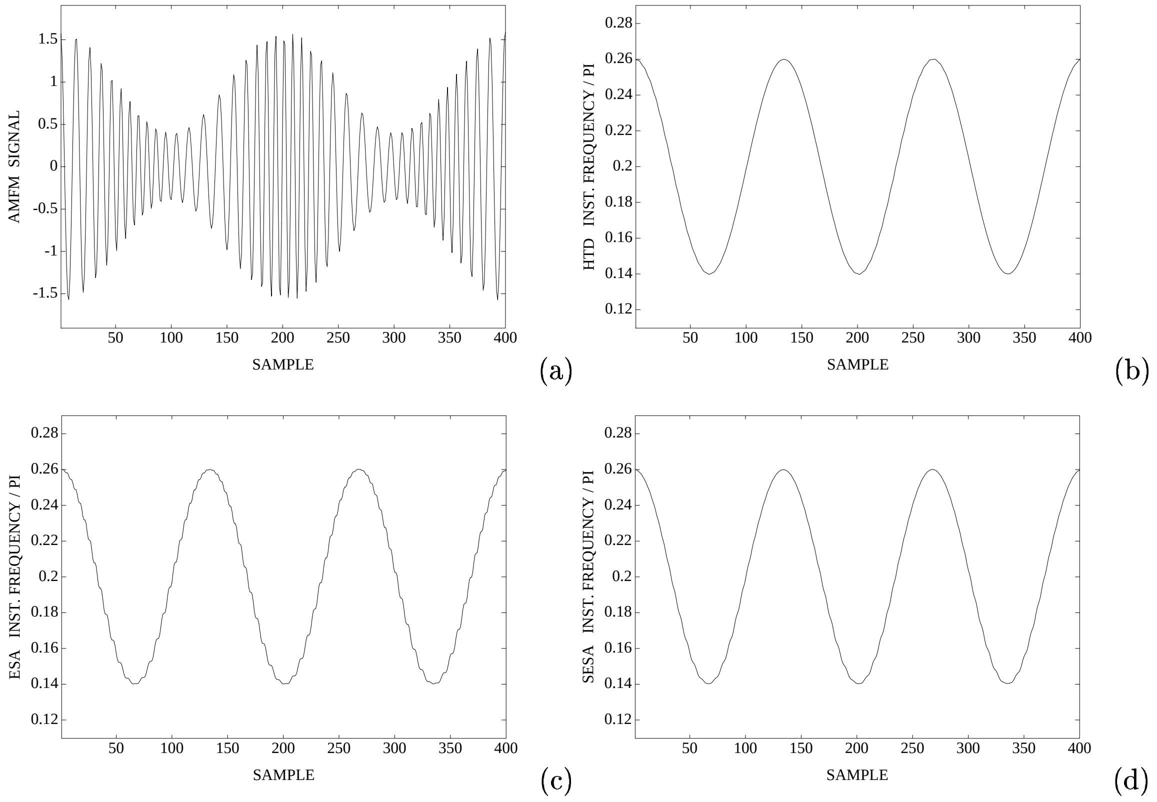


Figure 3.2: (a) AM–FM signal $x(n) = (1 + 0.6\cos[\pi n/100])\cos[\pi n/5 + 4\sin(3\pi n/200)]$. Estimated instantaneous frequency using the: (b) HTD, (c) ESA, and (d) SESA.

The corresponding sinusoidal amplitude and instantaneous frequency signals are:

$$a(n) \quad = \quad 1 + \kappa \cos(W_a\, n) \tag{3.20}$$

$$\Omega(n) \quad = \quad \Omega_c + \Omega_m \cos(W_f\, n) \tag{3.21}$$

The AM modulation index $\kappa \in (0,1)$ and the FM modulation depth $\Omega_m/\Omega_c \in (0,1)$ determine, respectively, the amount of AM and FM. W_a and W_f are the bandwidths of the amplitude and frequency modulating signals, respectively. An example of estimating the amplitude envelope and the instantaneous frequency using the long HTD (139–sample FIR design), the ESA and the SESA is shown in Figs 3.1 and 3.2 for an AM–FM/cosine signal with modulation amounts of 60% AM ($\kappa = 0.6$) and 30% FM ($\Omega_m/\Omega_c = 0.3$). We observe that all three algorithms yield good estimates for the amplitude envelope and instantaneous frequency signals. A few small ripples found in the ESA estimates are eliminated via the simple binomial smoothing involved in the SESA.

From extensive experiments on the class of AM–FM/cosine signals the performance of the demodulation algorithms was found to depend mainly upon the ratio of the carrier frequency over the bandwidth of the information signals $\Omega_c/W_{a,f}$, where $W_{a,f} = \max(W_a, W_f)$, the AM index κ, and the FM depth Ω_m/Ω_c. Henceforth in this section we assume that $W_{a,f} = W_a = W_f$. Extensive numerical comparisons among the HTD (using both the short 19–sample and the long 139–sample FIR implementations), the ESA and SESA on AM–FM/cosine signals were performed by varying $\Omega_c/W_{a,f}$, κ and Ω_m/Ω_c. A typical value for the ratio $\Omega_c/W_{a,f}$ for speech analysis applications is about 10, since an average formant value is at 2 kHz and the bandwidths of the amplitude/frequency modulating signals have been found at 200 Hz on the average. However, in communication systems the ratio $\Omega_c/W_{a,f}$ takes much higher values, e.g., in AM radio the ratio is in the order of 100, whereas in FM radio it is in the order of 1000. For our experiments, the case where $\Omega_c/W_{a,f} \approx 10$ will be referred to as *"speech specifications"* and the case where $\Omega_c/W_{a,f} \geq 100$ will be referred to as *"communications specifications"*.

Table 1 shows the mean absolute error for envelope and frequency estimation *averaged* over 100 different AM–FM/cosine signals with AM index varying from 5–50% (step 5%) and FM depth varying from 2–20% (step 2%). The carrier frequency was fixed at $\Omega_c = \pi/5$. The average errors are displayed for all four demodulation algorithms for $\Omega_c/W_{a,f} = 10$ and

TABLE 1: Percent Amplitude and Frequency Estimation Mean Absolute Errors using Demodulation Algorithms on AM–FM/cosine Signals ($\Omega_c = \pi/5$)

Algorithm	$\Omega_c/W_a = \Omega_c/W_f = 10$		$\Omega_c/W_a = \Omega_c/W_f = 100$	
	Amplitude error %	Frequency error %	Amplitude error %	Frequency error %
short HTD	4.41	4.60	4.39	4.46
long HTD	0.03	0.04	0.03	0.03
ESA	0.39	0.32	0.03	0.02
smooth ESA	0.25	0.22	0.02	0.01

TABLE 2: Percent Amplitude and Frequency Estimation Mean Absolute Errors using Demodulation Algorithms on AM–FM/cosine Signals with 30 dB Signal–to–Noise Ratio

Algorithm	$\Omega_c/W_a = \Omega_c/W_f = 10$		$\Omega_c/W_a = \Omega_c/W_f = 100$	
	Amplitude error %	Frequency error %	Amplitude error %	Frequency error %
short HTD	4.87	5.06	4.85	4.93
long HTD	1.83	2.15	1.84	2.19
ESA	4.81	4.43	4.71	4.35
smooth ESA	1.37	1.87	1.28	1.72

TABLE 3: Computational Complexity of Demodulation Algorithms. (Number of Operations Per Sample)

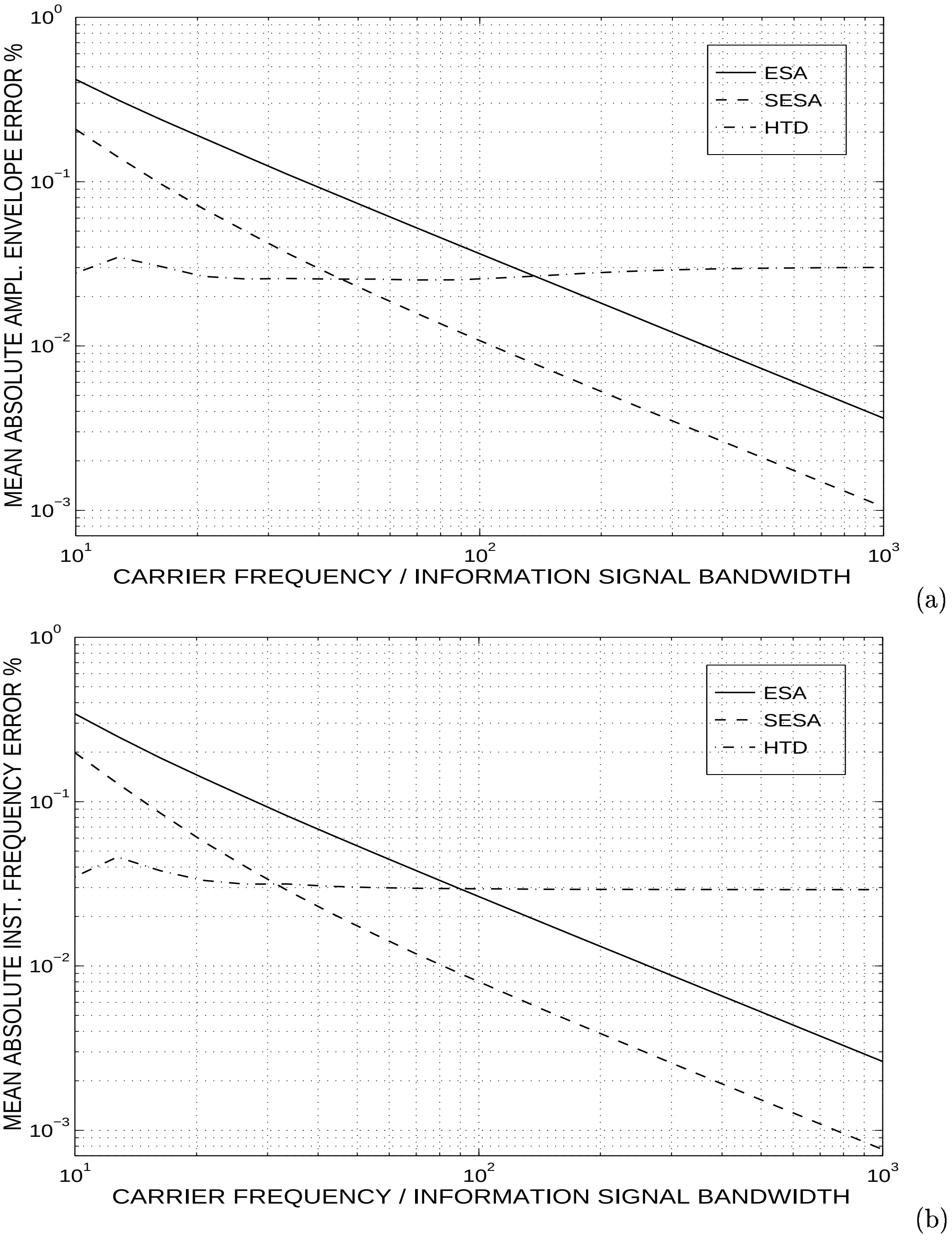
Algorithm	Additions	Multiplications	arccos$(\cdot)$	$\sqrt{(\cdot)}$	L
short HTD	12	8	1	1	20
long HTD	73	38	1	1	140
ESA	6	8	1	1	5
smooth ESA	24	8	1	1	11

* L is the number of samples in the moving window.

100. Overall, the long HTD yielded approximately one order of magnitude smaller error than the ESA and SESA for speech specifications. Yet, the short HTD (with approximately the same computational complexity as the ESA) performed much worse than the ESA. The smoothed ESA estimation error was 30–50% smaller than the error of the ESA without smoothing. For communications specifications, the errors of both the ESA and SESA were comparable or somewhat smaller to the long HTD error.

Next, we compare the performance of the demodulation algorithms in the presence of noise. Table 2 shows the mean absolute amplitude envelope and instantaneous frequency estimation error averaged over the 100 cases of AM–FM/cosine signals used for Table 1, in the presence of added white Gaussian noise at a signal–to–noise ratio (SNR) of 30 dB. For both speech and communications specifications, the long HTD performed better than the ESA. This is expected, since the HTD involves an integral transform that does implicit smoothing, whereas the ESA involves a an "instantaneous" differential operator. Interestingly, the smoothed ESA, which uses the simple 7–point binomial smoother, yielded an error comparable or smaller to the long HTD error. Finally, we note that the comparison of the demodulation algorithms are based on numerical simulations and refer to the special case of AM–FM/cosine signals, for both the noise–free and the noise–corrupted case. A theoretical analysis of the HTD approach for random signals can be found in [81], and a theoretical analysis for the performance of the ESA in the presence of Gaussian noise has been recently developed in [8, 9].

Table 3 shows that out of the four algorithms the ESA has the smallest computational complexity and needs the smallest number L of input samples per single output estimate in its moving window. The smoothed ESA needs a few more additions for the binomial smoothing and a window twice as long. The short HTD has similar computational complexity to the ESA but needs a four times longer window. Finally, the long HTD has about one order of magnitude higher computational complexity than the ESA and 3–4 times higher than the SESA. The biggest drawback of the long HTD is that it requires a window more than one order of magnitude wider than the window of the ESA and SESA. Hence, the ESA and SESA have the advantage of adapting instantaneously and needing an extremely small number of input samples to operate. Analysis of speech signals using



(a)

(b)

Figure 3.3: (a) ESA, SESA and HTD mean absolute envelope estimation error (%) for $\Omega_c/W_{a,f} \in [10, 1000]$ ($\Omega_c = \pi/5$ is the carrier frequency and $W_{a,f}$ is the bandwidth of the AM and FM modulating signals.) (b) ESA, SESA and HTD mean absolute frequency estimation error (%). Each point in these curves is the average error over 100 experiments on AM–FM/cosine signals as the percent of AM varies from 5–50% and of FM from 2–20%.

one of the demodulation algorithms is done on a short–time basis. From speech analysis experiments, we observed that the long HTD implementation needs an FIR length $L \approx N$, where N is the average length of the short–time speech analysis frame. Hence, in this case, the computational complexity of the HTD is quadratic $O(N^2)$. In contrast, the complexity of the ESA and SESA is always linear $O(N)$, and the multiplicative constant is very small.

Finally, in Fig. 3.3 we compare the amplitude envelope and instantaneous frequency estimation errors for the three algorithms (ESA, SESA and long HTD), for values of the ratio $\Omega_c/W_{a,f}$ ranging from 10 to 1000 (Ω_c is set to $\pi/5$ throughout the experiment). Each point in the error curves represents the average error over 100 experiments for AM–FM/cosine signals with a varying AM index $\kappa \in [0.05, 0.5]$ and a varying FM depth $\Omega_m/\Omega_c \in [0.02, 0.2]$ (as in Table 1). We observe that the ESA error is decreasing linearly with the ratio $\Omega_c/W_{a,f}$, while the HTD error remains approximately constant. For $\Omega_c/W_{a,f} = 10$ the ESA error is one order of magnitude larger than the HTD error. As the ratio approaches 100 the ESA and HTD error are of comparable magnitude. Finally, when the ratio is in the neighborhood of 1000 the ESA error is one order of magnitude smaller. The SESA error decreases linearly with $\Omega_c/W_{a,f}$ and is approximately half of the ESA error. As the bandwidth of the information signals decreases the performance of the SESA relative to the ESA improves.

3.4 Experiments on Real Speech

In this section, the above presented demodulation algorithms are applied to real and synthetic speech signals to track the envelope and the frequency modulation of speech resonances. A single speech resonance is modeled as an exponentially damped AM–FM signal (see Eq. (1.11)); the speech signal is the sum of such AM–FM signals.

Before applying the demodulation algorithms to a speech resonance, we must first extract the resonance through bandpass filtering. For this purpose, we use a Gabor filter because it is optimally compact and smooth both in the time and frequency domain. The impulse and frequency response of the real Gabor filter is

$$\begin{aligned} h(t) &= \exp(-\alpha^2 t^2) \cos[2\pi\nu t] && (3.22) \\ H(w) &= \frac{\sqrt{\pi}}{2\alpha} \left(\exp\left[-\frac{\pi^2 (w-\nu)^2}{\alpha^2} \right] + \exp\left[-\frac{\pi^2 (w+\nu)^2}{\alpha^2} \right] \right) && (3.23) \end{aligned}$$

The frequency ν of the filter is chosen equal to the center formant frequency F. The parameter α controls the bandwidth of the filter. The effective RMS bandwidth of the Gabor filter is equal to $\alpha/\sqrt{2\pi}$, i.e., $\sqrt{2\pi}$ times the RMS bandwidth [22]. In this case, the effective RMS bandwidth is comparable to the 3 dB bandwidth. Note that in discrete time the impulse response $h(n)$ is a sampled and truncated version of Eq. (3.22).

In addition to the need of bandpass filtering to extract a single resonance, voiced speech signals have also a specific quasi–periodic structure. Henceforth, we examine the performance of the long HTD and the SESA for speech resonance demodulation. In particular, we investigate how the demodulation algorithms are affected by the pitch periodicity, by bandpass filtering and by phonemic transitions.

3.4.1 Effects of Pitch

The effects of pitch on envelope and frequency estimation are studied first on a synthetic resonance signal $s(n)$. In this example, $r(n)$ is the output of a linear time–invariant speech resonator with a single resonance at 1300 Hz (3 dB bandwidth = 30 Hz), excited by a periodic sequence of unit pulses with fundamental frequency at 100 Hz. Fig. 3.4 (a), (b) show $r(n)$ and the excitation signal. The amplitude envelope and instantaneous frequency estimates for the energy operator and the Hilbert transform demodulation algorithms are shown in Fig. 3.4 (c)–(f). The HTD envelope estimate (c) shows clearly the exponential decay of the actual envelope, yet it also displays misleading modulations around the instants that the pitch pulses occur. Similarly, the HTD instantaneous frequency estimate (d) tracks correctly the formant frequency at 1300 Hz, but in the neighborhood of the pitch pulses the frequency estimate is heavily modulated. The SESA estimated envelope (e) consists of decaying exponentials interrupted by a small spike at each pitch pulse. In the same way, the instantaneous frequency SESA estimate (f) is constant everywhere at 1300 Hz, except at the location of the pitch pulses where large double spikes occur.

In brief, the SESA envelope and instantaneous frequency estimation error is concentrated at the excitation instants, while for the HTD the estimation error is significant in a time interval of 3–5 msecs around each excitation pulse. Clearly, the HTD by using an integral transform does implicit smoothing (lowpass filtering) to the information signals.

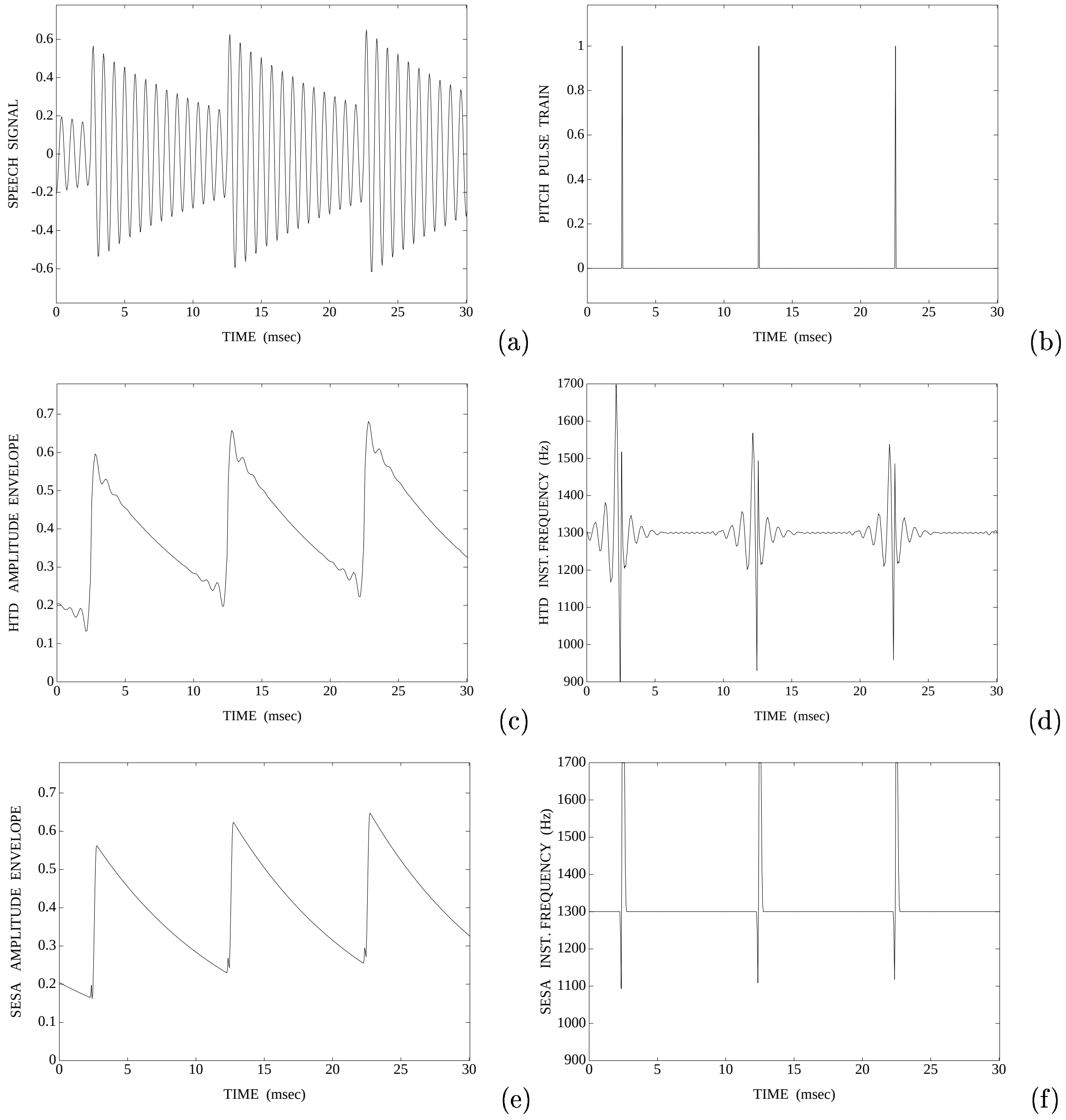


Figure 3.4: (a) Synthetic speech signal with a single formant at 1300 Hz and fundamental frequency at 100 Hz (sampling frequency at 20 kHz). (b) The excitation signal is a sequence of pulses with a 10 msecs period. (c) Estimated amplitude envelope using the long HTD. (d) Estimated instantaneous frequency using the long HTD. (e) Estimated amplitude envelope using the SESA. (f) Estimated instantaneous frequency using the SESA.

Thus, the high frequency component of the excitation event is filtered out and what we observe is a lowpass filtered spike ("ringing"). On the other hand, the energy operator is an instantaneous differential operator, whose discrete implementation involves a very short analysis window (a very few input samples per output sample). As a result, the SESA estimates have superior time resolution compared to the HTD ones, e.g., abrupt transitions are better preserved.

Of interest is also the general case of a discontinuity at the envelope or the instantaneous frequency of an AM–FM speech–like signal. Consider, for example, the case where a jump occurs to the carrier frequency f_c, causing the AM–FM signal to be discontinuous. As in the previous example, the HTD estimation breaks down in the neighborhood of the instant when the discontinuity occurs, presenting erroneous modulations, especially, in the instantaneous frequency estimate. The SESA estimates, however, have a (single or double) spike at the jump instant and are "correct" elsewhere. As a result, the actual envelope and instantaneous frequency around the discontinuity are easily recoverable from the SESA followed by median filtering. In addition, the "spiky" nature of the energy signals at the instant of the jump informs us of the transition/event. Thus, the energy operator can serve as an event detector, as opposed to the Hilbert transform which tends to smooth out discontinuities. An application of the energy operator event detector property in underwater acoustics can be found in [89].

An example of the effects of a carrier frequency discontinuity on the instantaneous frequency estimate is shown in Fig 3.5. The carrier frequency of an AM–FM signal $x(n)$ jumps by 1.43% at the 250th sample, causing a discontinuity to the signal itself. The HTD and SESA frequency estimates are shown in (c), (d). Clearly, the HTD frequency estimate presents erroneous modulations in the neighborhood of the jump, while the SESA error is concentrated only in 10–12 samples, around the point of discontinuity. In general, since the "long HTD" analysis window is an order of magnitude longer that the SESA window (see Table 3), the estimation error caused by a signal discontinuity is spread over a longer time segment for the HTD than for the ESA.

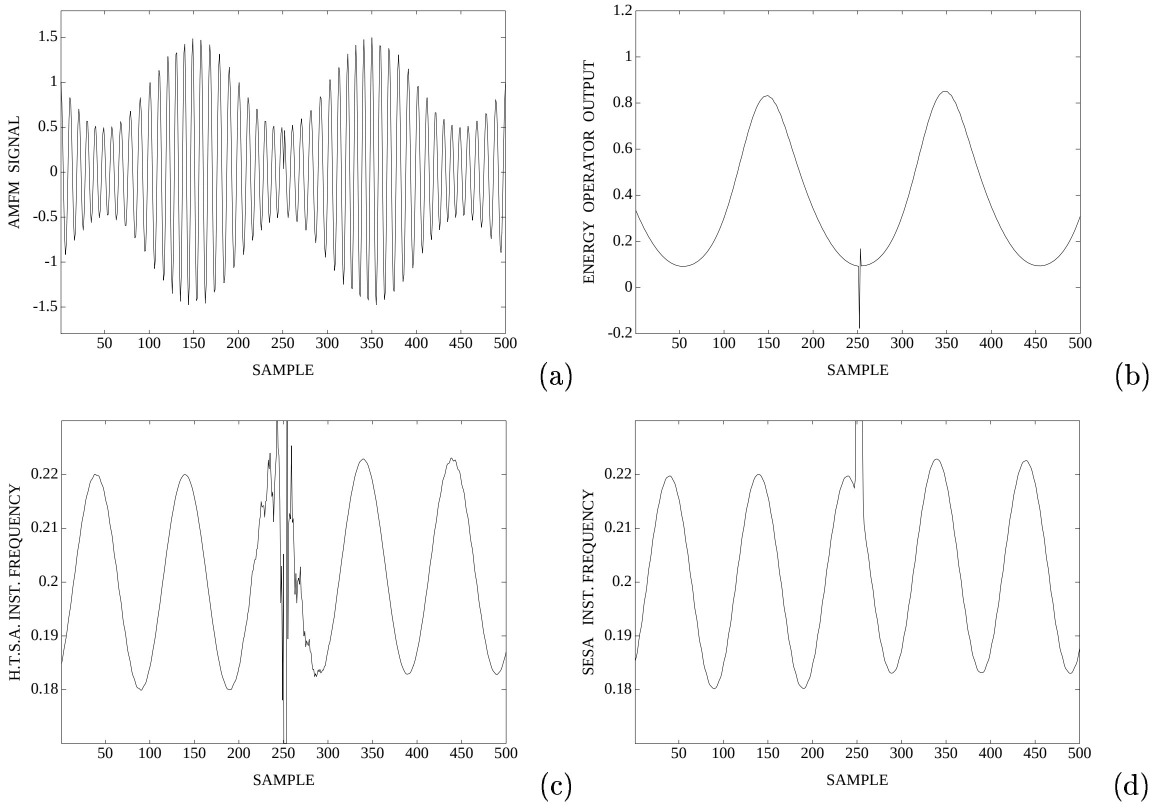


Figure 3.5: (a) The AM–FM signal $x(n) = (1 + 0.5\cos[\pi n/100])\cos[\mu\pi n/5 + 2\sin(\pi n/50) + \phi]$, where $\mu = 1$ for the first 250 samples and $\mu = 1.0143$ for the rest. (b) The Energy Operator output $\Psi[x(n)]$. Estimated instantaneous frequency of $x(n)$ using: (c) long HTD, (d) SESA.

3.4.2 Effects of Bandpass Filtering

For any AM–FM signal $x(t) = a(t)\cos[\phi(t)]$ we may write:

$$x(t) = a(t)\cos[2\pi f_c t + p(t)] = a_1(t)\cos[2\pi f_c t] - a_2(t)\sin[2\pi f_c t] \tag{3.24}$$

where

$$a(t) = \sqrt{a_1^2(t) + a_2^2(t)}\,, \qquad p(t) = \arctan\left[\frac{a_2(t)}{a_1(t)}\right]. \tag{3.25}$$

Next, we filter $x(t)$ through a bandpass filter with impulse response

$$h(t) = h_\ell(t)\cos[2\pi f_c t] \tag{3.26}$$

where $h_\ell(t)$ is the impulse response of the corresponding lowpass filter and f_c is the carrier frequency of the AM–FM signal $x(t)$. The filtered signal $\tilde{x}(t) = x(t) * h(t)$ is approximately equal to (the conditions under which the approximation holds can be found in [80, ch. 7])

$$\tilde{x}(t) = \tilde{a}(t)\cos[2\pi f_c t + \tilde{p}(t)] \approx \frac{1}{2}\,[a_1(t) * h_\ell(t)]\cos[2\pi f_c t] - \frac{1}{2}\,[a_2(t) * h_\ell(t)]\sin[2\pi f_c t] \tag{3.27}$$

and the new amplitude envelope and instantaneous frequency are

$$\tilde{a}(t) \approx \frac{1}{2}\,\sqrt{[(a(t)\cos[p(t)]) * h_\ell(t)]^2 + [(a(t)\sin[p(t)]) * h_\ell(t)]^2} \tag{3.28}$$

$$\tilde{f}_i(t) = f_c + \frac{1}{2\pi}\frac{d}{dt}[\tilde{p}(t)] \approx f_c + \frac{1}{2\pi}\frac{d}{dt}\,\arctan\left[\frac{(a(t)\sin[p(t)]) * h_\ell(t)}{(a(t)\cos[p(t)]) * h_\ell(t)}\right]. \tag{3.29}$$

We conclude that the amplitude envelope and instantaneous frequency of a bandpass filtered AM–FM signal $\tilde{x}(t)$ are lowpass filtered versions of the actual information signals. Indeed, for a slowly–varying phase modulating signal $p(t)$ the amplitude envelope of the bandpass filtered signal is simply

$$\tilde{a}(t) \approx a(t) * h_\ell(t). \tag{3.30}$$

Effects of Gabor bandpass filtering

An example of how bandpass filtering affects the SESA and HTD estimation is shown in Fig. 3.6. The vowel /ah/ with formants at 600, 1200, 2500 and 3600 Hz (and 3 dB bandwidth at 30, 50, 80 and 110 Hz, respectively) is synthesized using time–invariant linear resonators in cascade, excited by a sequence of unit pulses with fundamental frequency at 100 Hz. Then,

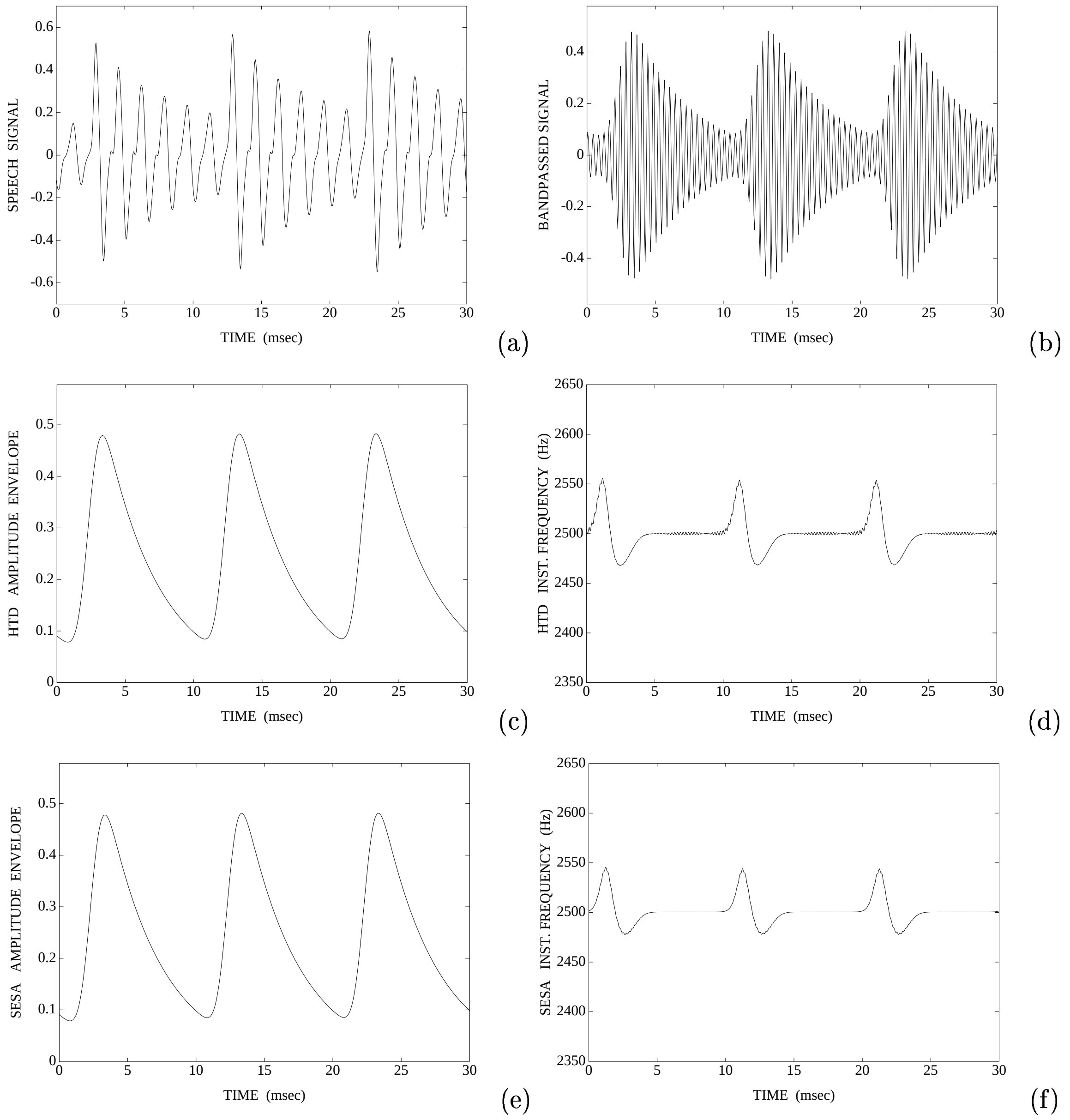


Figure 3.6: (a) Synthetic speech signal $s(n)$ (vowel /ah/) with formants at 600, 1200, 2500 and 3600 Hz and fundamental frequency at 100 Hz (sampling frequency at 20 kHz). (b) Speech signal after Gabor filtering around the formant at $\nu = 2500$ Hz (bandwidth of the filter is 400 Hz). (c), (d) Estimated amplitude envelope and instantaneous frequency using the long HTD. (e), (f) Estimated amplitude envelope and instantaneous frequency using the SESA.

the synthetic signal is bandpass filtered around its third formant, using a Gabor filter with center frequency at $\nu = F_3 = 2500$ Hz and effective RMS bandwidth of 400 Hz. In Fig. 3.6 (a), (b), three pitch periods of the synthetic vowel /ah/ and the extracted resonance are shown. The HTD and SESA estimates for the amplitude envelope and the instantaneous frequency of the resonance at 2500 Hz are shown in (c)–(f). Interestingly, both demodulation algorithms (HTD, SESA) produce almost identical estimates. The amplitude envelope estimates (c), (e) are exponentially decaying with smooth transitions at the instants when pitch pulses occur. The frequency estimates (d), (f) are everywhere equal to the center formant frequency ($F_3 = 2500$ Hz), apart from a 3–5 msecs segment around the pitch pulses, where they deviate considerably from F_3.

In Fig. 3.4, we have displayed the amplitude envelope and instantaneous frequency estimates, when filtering was not needed to extract the resonance (single formant case). We observe that by bandpass filtering the original signal, double spikes turn into smooth "sinusoidal" curves (SESA frequency estimate (f)) and jumps into smooth transitions (SESA envelope estimate (e)). This is anticipated, because bandpass filtering eliminates much of the higher frequency components of the amplitude envelope and instantaneous frequency signals (see Eqs. (3.28), (3.29)). In that sense, after bandpass filtering of the original signal, the SESA and HTD display similar results, as now both algorithms involve smoothing (or lowpass filtering) of the information signals. When bandpass filtering is applied, the excellent time–resolution of the SESA is blurred (and the event–detector property is somewhat lost).[2] In brief, the effect of the Gabor filter is to smooth the spikes and the abrupt jumps (if any) of the original estimates, especially for the ESA where high frequency components are preserved.

The Gabor bandpass filtering has similar effects on the amplitude envelope and instantaneous frequency signals of real speech resonances. The information signals, though, are different from the linear synthetic case. As shown in Fig. 1.3 the linear model can capture only some of the rich structure of the information signals of the resonance. The actual amplitude envelope and instantaneous frequency are in many cases heavily modulated, es-

[2] Even in this case, the SESA has better time–resolution because the analysis window used by the combination of filtering and the SESA is approximately half of the size of the window that the combination of filtering and HTD uses.

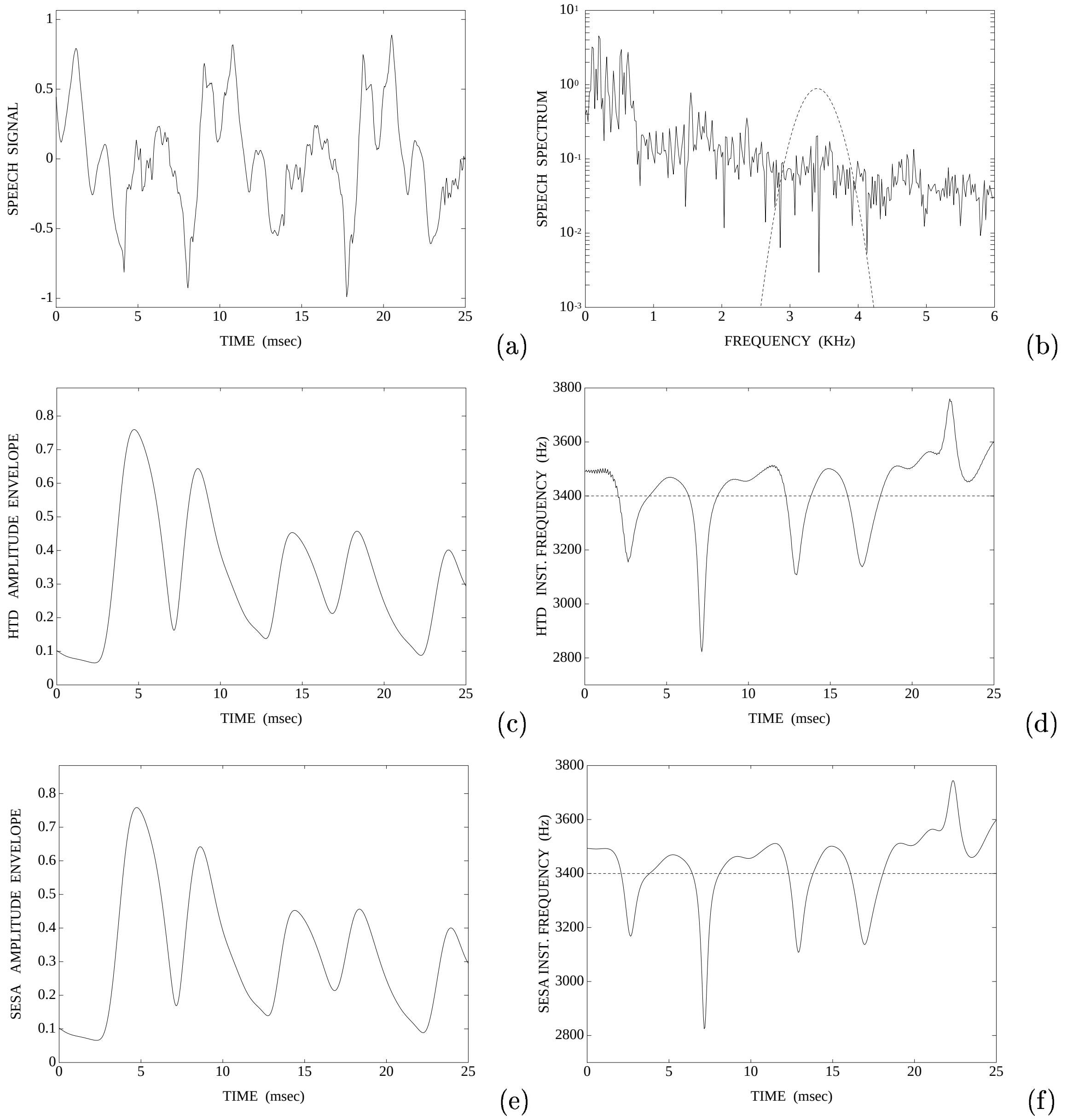


Figure 3.7: (a) Speech signal $s(n)$ (vowel /ih/ sampled at 20 kHz). (b) Spectral magnitude of the speech signal $s(n)$ and of the Gabor filter ($\nu = 3400$ Hz, bandwidth 400 Hz). (c) Estimated amplitude envelope using the long HTD, after Gabor filtering of $s(n)$ around the spectral peak at $F = 3400$ Hz. (d) Estimated instantaneous frequency using the long HTD (the dashed line shows the center frequency of the Gabor filter). (e) Estimated amplitude envelope using the SESA. (f) Estimated instantaneous frequency using the SESA.

pecially for higher formants. In Fig. 3.7, we present a real speech example of resonance demodulation (vowel /ih/) for the formant at $F \approx 3400$ Hz. A Gabor filter with center frequency $\nu = 3400$ Hz and 400 Hz bandwidth was used to extract the resonance. The HTD and SESA estimated envelope are shown in (c), (e) respectively. The estimates present only minor differences at envelope minima and apart from that they are almost identical. The instantaneous frequency estimates (d), (f) are also very similar. Note the rich amplitude and frequency modulations in the speech resonance in question.

In numerous examples of speech analysis using the SESA and the HTD, we saw only minor differences in the estimated amplitude envelope and instantaneous frequency contours. In some cases, the Hilbert transform algorithm seems to yield slightly smoother estimates than the SESA, especially for frequency estimation of low formants. Also, in a few isolated instances, the SESA may produce narrow spikes, e.g., at envelope minima and at the corresponding places at the instantaneous frequency estimate. Note that the minor differences between the ESA and HTD estimators usually occur around the envelope minima. Overall, both algorithms yield similar and equally satisfying results for real speech analysis. However, the SESA is faster and uses an shorter analysis window.

On determining the Gabor filter parameters

A question that arises in speech demodulation experiments, is what happens to the envelope and the instantaneous frequency estimates, when the center frequency of the Gabor filter is not exactly equal to the center frequency of the formant. Using AM–FM speech–like signals, we observed that the estimated amplitude envelope and instantaneous frequency are close to the actual ones, for center frequency differences less than 100 Hz. In real speech experiments, shifting the center frequency of the Gabor filter in the neighborhood of a formant affects the envelope and frequency contours mainly around envelope minima. Specifically, the instantaneous frequency seems to be unstable around these points, presenting large peaks or valleys. In order to avoid such instabilities, the center formant frequency must be determined accurately (see Chapter 4).

Determining the bandwidth of the Gabor filter is more difficult; an obvious choice is the bandwidth of the signal in question. Carson's rule determines the approximate bandwidth of an AM–FM signal to be twice the sum of the maximum frequency deviation plus the bandwidth of the AM and FM information signals. For our experiments, this would correspond to a Gabor filter effective RMS bandwidth in the range of 1200–2000 Hz. In real speech, we need to isolate (via filtering) spectral peaks that are 500–1000 Hz apart. Thus, in order to avoid the effects of the neighboring formants (see Section 3.4.3), we must limit the Gabor filter bandwidth to more conservative values, e.g. 400 Hz. Next, we consider how the envelope and instantaneous frequency estimates contours are affected when the bandwidth of the filter is smaller than the effective bandwidth of the signal.

We used synthetic AM–FM speech–like signals to address this question. The frequency modulation depth Ω_m/Ω_c parameter was selected to be 10% (in speech analysis rarely have we found larger amounts of FM). For various values of AM, we found that filter bandwidths in the range 400–600 Hz give good envelope and frequency estimates, i.e, correct shape and maximum absolute error from 5%–10%.

In real speech, it is hard to determine how the Gabor filter bandwidth affects the envelope and instantaneous frequency because of the effects of the neighboring formants. It is clear though, that for smaller filter bandwidths the bandwidth of the estimated information signals decreases.

3.4.3 Effects of Neighboring Spectral Peaks

A neighboring spectral peak that has not been thoroughly eliminated through bandpass filtering can seriously affect the estimated envelope and instantaneous frequency contours of the current formant. In [54], a model has been proposed for dealing with this problem. Specifically, suppose that f_c and f_x are the center frequencies of a formant and its neighbor. Then the bandpass filtered signal $y(t)$ for $\lambda \ll 1$ (where λ is the relative gain of the neighboring formant vs. the center formant) is

$$\begin{aligned} y(t) &= \cos(2\pi f_c t) + \lambda \cos(2\pi f_x t + \theta) \\ &\approx \cos[2\pi f_c t - \lambda \sin(2\pi f_\omega t - \theta)] + \lambda \cos(2\pi f_\omega t - \theta) \cos(2\pi f_c t) \end{aligned} \tag{3.31}$$

where $f_\omega = f_c - f_x$. Thus, the neighboring spectral peak modulates the envelope and instantaneous frequency estimates, with a modulation frequency f_ω equal to the difference of the central formant frequencies of the two spectral peaks.

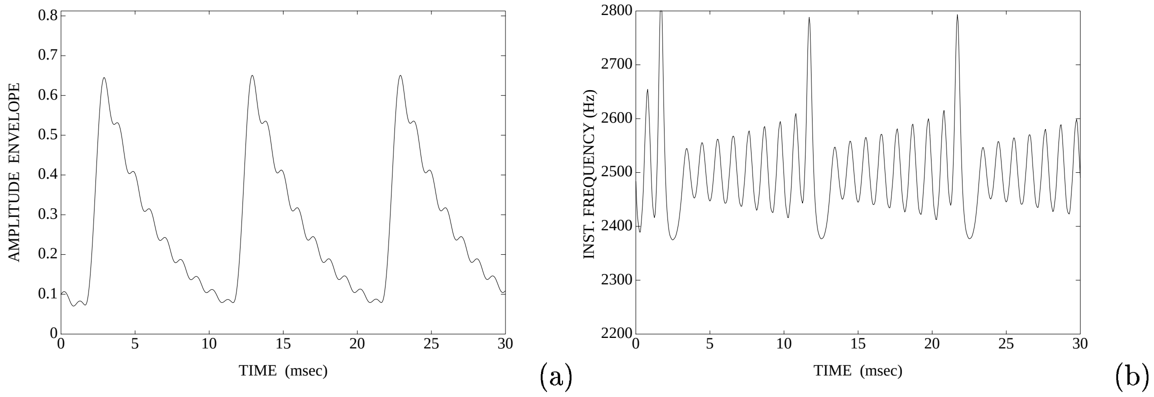


Figure 3.8: (a) SESA estimated amplitude envelope for the third formant ($F_3 = 2500$ Hz) of the bandpassed synthetic speech vowel $s(n)$ (formants at 550, 1550, 2500 Hz and fundamental frequency at 100 Hz, Gabor center frequency $\nu = 2500$ Hz and bandwidth 620 Hz). (b) Estimated instantaneous frequency using the SESA.

In Fig. 3.8, we present experimental evidence in support of this model. The third resonance of a synthetic speech vowel with formants at 550, 1550 and 2500 Hz and fundamental frequency of 100 Hz is analyzed. We have displayed in Fig. 3.6, the estimated amplitude envelope and instantaneous frequency contours when the neighboring formants have been thoroughly eliminated by a Gabor filter of the appropriate bandwidth. In this case, the filter is centered at $F_3 = 2500$ Hz and the bandwidth of the filter is chosen to be 620 Hz so that the formant at $F_2 = 1550$ Hz is still “in play.” The estimates for the amplitude envelope and the instantaneous frequency of F3 are displayed in Fig. 3.8 (a) and (b). The estimates are modulated, with a modulation frequency equal to the difference between the two formant frequencies, i.e. $F_3 - F_2 = 950$ Hz. The amplitude of the modulations increases as the Gabor filter bandwidth increases. Similarly, we have observed modulations in the amplitude envelope and the instantaneous frequency estimates of real speech resonances due to interaction with neighboring formants.

3.4.4 Analysis of Transitions in Real Speech

Formant finding and feature extraction during speech transitions is not a simple task, because most speech parameters change rapidly and short–time analysis assumptions do not

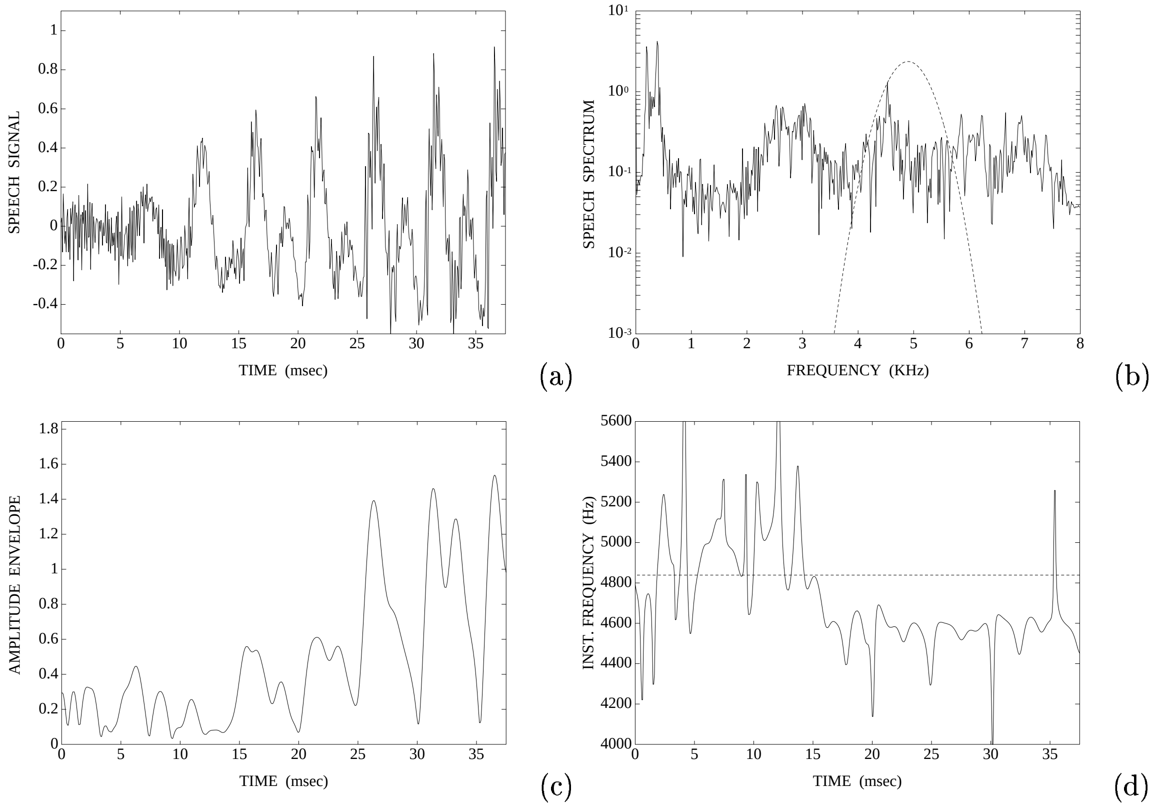


Figure 3.9: (a) Speech signal: transition between unvoiced consonant /s/ and vowel /eh/, sampled at 16 kHz. (b) Spectral magnitude of the speech signal and of the Gabor filter (ν = 4850, bandwidth = 600 Hz). (c) Estimated amplitude envelope using SESA, after Gabor filtering around ν = 4850 Hz. (d) Estimated instantaneous frequency (the dashed line shows the center frequency of the Gabor filter).

hold. The usual approach to this problem is to use continuity and smoothness constraints to interpolate the estimated quantities from neighboring short speech segments. Unfortunately, this method does not always produce satisfactory results.

We know that the energy separation algorithm has better time resolution than conventional short–time analysis methods since the estimates are computed from an extremely short speech segment. Also, the Gabor filter followed by the SESA introduces minimal blurring/smoothing of rapidly varying modulating signals of the speech resonances. Thus, one can use the amplitude envelope and instantaneous frequency estimates for parameter estimation and feature extraction during transitions in real speech.

An example of tracking a transient formant frequency is presented in Fig. 3.9, during a transition from the unvoiced consonant /s/ to the vowel /eh/. In this example, the vowel /eh/ has a strong spectral peak at 4500 Hz, while the consonant /s/ does not have a formant around this frequency. The speech signal is bandpass filtered using a Gabor filter with center frequency at $\nu = 4850$ Hz and bandwidth 600 Hz, in order to follow the formant track during the transition from /s/ to /eh/. The SESA amplitude envelope estimate (c) displays the energy increase as we pass from the unvoiced to the voiced sound. Similar amplitude modulation patterns can be seen for both sounds. The instantaneous frequency estimate (d) shows the center frequency of the formant changing rapidly during the transition. We observe a 500 Hz shift in the center formant frequency in a time period of 5–10 msecs. Note also that the vowel reaches a steady state 10–15 msecs after voicing begins.

3.5 Conclusions

In this chapter, we have compared two different approaches for estimating the time–varying amplitude envelope and instantaneous frequency of general AM–FM signals, as well as of speech resonances: the energy separation algorithms (ESA, SESA) involving a nonlinear differential operator and the Hilbert transform demodulation algorithm (HTD) involving a linear integral transform. We have compared the demodulation algorithms first on general synthetic AM–FM signals and then on speech vowel resonance signals extracted via Gabor bandpass filtering. Next, some important issues related to the application of the amplitude/frequency separation algorithms to speech resonance demodulation were investigated and discussed. These include choosing the Gabor bandpass filter parameters, the effect of neighboring formants, and transient formant analysis.

After extensive experiments on synthetic AM–FM signals, we have found that, in the absence of noise, when the ratio R of the carrier frequency over the information signals bandwidth is in the order of 10 (as in speech applications) the ESA yields a mean absolute error in the order of $10^{-1}\%$; when this ratio becomes 100 or 1000 (as in communication applications) the ESA yields errors in the order of $10^{-2}\%$ and $10^{-3}\%$, respectively. Note that even in the worst case ($R = 10$) the ESA yields a relatively small error. The SESA almost always reduces the ESA error by about 50%, except for very small values of R. The HTD yields an error in the order of $10^{-2}\%$ for all the above values of R. In the presence of 30 dB noise, both HTD and SESA yield errors in the order of 1%, with the SESA yielding the smallest error. In the analysis of short time segments of speech vowel signals (synthetic or real) bandpass filtered around their formants, the SESA was found to yield modulating signals very close to the ones obtained via the HTD.[3] The fact that both algorithms yield similar results for speech signals is due to the bandpass filtering, which blurs the instantaneously varying features of the time waveform.

For all signals, both ESA and SESA have very small computational complexity, linear in the number of input samples. The HTD has about one order of magnitude higher complexity. For speech applications the HTD complexity becomes quadratic in the number of samples. Finally, while the ESA or SESA requires for its operation an extremely small number of input samples in its moving window, the HTD requires an order of magnitude longer window.

In short, the SESA was found to yield comparable estimation errors to the HTD for tracking AM–FM modulations in speech signals. For communications applications, the SESA yields a smaller error. In addition, the SESA has the advantages over the HTD of smaller computational complexity and faster time–adaptivity. Finally, the use of an energy operator gives the SESA an additional interesting intuitive feature: the tracking of the energy required for generating the AM–FM signal and the separation of the energy into amplitude and frequency components.

[3] The accuracy of the HTD estimates was superior for formants of very low frequency (below approximately 500 Hz).

Chapter 4

Speech Analysis I: Formant Tracking

4.1 Introduction

Formant tracking is a interesting and challenging speech analysis problem because the formant locations are very important cues for both human and machine speech recognition. Formant trajectories have been also used successfully both in speech coding and speech synthesis applications. During the past four decades many different formant tracking schemes have been proposed. Currently, most of the widely used formant tracking algorithms are based on linear prediction (LP) analysis [65, 16]. LP formant trackers encounter problems with nasal formants, spectral zeros, and bandwidth estimation. These deficiencies stem from the fact that LP is a parametric method that does not model spectral valleys. In addition, LP is a linear model unable to adequately model speech acoustics. One can overcome some of these deficiencies by using a pole–zero model for formant tracking [105]. Other more complex formant tracking algorithms use the extended Kalman filter [73] or hidden Markov models [47]. Alternatively, we propose here a multiband filtering energy demodulation approach to speech analysis in the framework of the AM–FM modulation model that is easy to implement and overcomes most of the deficiencies of LP.

The AM–FM modulation model (introduced in Section 1.4) represents a single speech resonance as a signal $r(t)$ with combined amplitude and frequency modulation

$$r(t) = a(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \tag{4.1}$$

where $f_c \triangleq F$ is the "center value" of the formant frequency, $q(t)$ is the frequency modulating signal, and $a(t)$ is the time–varying amplitude. The instantaneous formant frequency signal is defined as $f(t) = f_c + q(t)$. Finally, speech $s(t)$ is modeled as the sum $s(t) = \sum_{k=1}^{N} r_k(t)$ of N such AM–FM signals, one for each formant. As discussed in Chapter 3, the amplitude envelope $|a(t)|$ and the instantaneous frequency $f(t)$ signals are obtained by applying a demodulation algorithm on the speech resonance $r(t)$. In addition, bandpass filtering is used to obtain a single speech resonance before demodulation can be performed. These two steps of speech analysis in the framework of the AM–FM modulation model will be referred to as *multiband demodulation analysis* (MDA) [9].

In this chapter, we combine the amplitude envelope $|a(t)|$ and the instantaneous frequency $f(t)$ signals of a resonance $r(t)$ into formant frequency and bandwidth estimates. We propose two short–time frequency measures for estimating the average frequency of a speech (frequency) band: the *mean instantaneous frequency*, which has been used for formant tracking in [26] and the *mean amplitude weighted instantaneous frequency*, a time domain equivalent of the first central spectral moment [15]. Based on the weighted frequency estimate, the modulation model, and the multiband filtering/demodulation scheme, we propose the *multiband demodulation formant tracking algorithm.* The algorithm produces reliable formant tracks and realistic formant bandwidth estimates. In addition, the MDA approach to formant tracking is non–parametric, easy to implement, and avoids most of the drawbacks of LP–based formant trackers.

The organization of this chapter is as follows. First, the analysis tools of the modulation model are reviewed, i.e., multiband filtering and demodulation. Next, we evaluate the performance of the (unweighted and weighted) short–time formant frequency and bandwidth estimates on both synthetic signals and speech. The multiband formant tracking algorithm is introduced in Section 4.3. There the speech signal is analyzed through a bank of real Gabor filters with fixed center frequencies, demodulated, and short–time frequency and bandwidth estimates are computed for each band. A simple decision algorithm converts the short–time estimates to raw formants and ultimately to formant tracks. Finally, in

Section 4.6 performance and implementation issues are discussed.

4.2 Multiband Filtering and Demodulation

A speech resonance is extracted from the speech signal through filtering using a real Gabor bandpass filter with impulse response $h(t)$ and frequency response $H(w)$

$$\begin{aligned} h(t) &= \exp(-\alpha^2 t^2)\cos(2\pi\nu t) & (4.2) \\ H(w) &= \frac{\sqrt{\pi}}{2\alpha}\left(\exp\left[-\frac{\pi^2(w-\nu)^2}{\alpha^2}\right] + \exp\left[-\frac{\pi^2(w+\nu)^2}{\alpha^2}\right]\right) & (4.3) \end{aligned}$$

where ν is the center frequency of the filter chosen equal to the center formant frequency and α is the bandwidth parameter. The effective RMS bandwidth of the filter is equal to $\alpha/\sqrt{2\pi}$.[1]

Although bandpass filters with an abrupt frequency cutoff are typically used in most analysis/synthesis systems, we find that the Gabor filter by being optimally compact and smooth both in the time and frequency domain provides accurate amplitude and frequency estimates in the demodulation stage that follows. In [9], one can find a detailed discussion on the advantages of Gabor wavelets for multiband energy demodulation.

The *energy amplitude/frequency separation algorithm* (ESA) was introduced in Section 2.1.1 to demodulate a speech resonance $r(t)$ into amplitude envelope $|a_E(t)|$ and instantaneous frequency $f_E(t)$ signals. The ESA is simple, computationally efficient, and has excellent time resolution. An alternative way to obtain $|a_H(t)|$ and $f_H(t)$ estimates is through the Hilbert transform demodulation (HTD) (see Section 3.2.2). As we have seen in Chapter 3 the HTD and the ESA produce similar results for speech resonance demodulation, but the HTD has higher computational complexity. In addition, the performance of both the HTD and especially the ESA is poor for a low first formant frequency. When the first formant frequency is close to the fundamental frequency the HTD provides smoother estimates for the first formant amplitude and frequency signals. The HTD will be used occasionally in this chapter.

[1] The effective RMS bandwidth is defined in [22] as $\sqrt{2\pi}$ times the RMS bandwidth.

4.3 Formant Frequency and Bandwidth Short–Time Estimates

Simple short–time estimates for the frequency F and bandwidth B of a formant candidate, respectively, are the *unweighted* mean F_u and standard deviation B_u of the instantaneous frequency signal $f(t)$, i.e.,

$$F_u = \tfrac{1}{T} \int_{t_0}^{t_0+T} f(t)\, dt \tag{4.4}$$

$$[B_u]^2 = \tfrac{1}{T} \int_{t_0}^{t_0+T} (f(t) - F_u)^2\, dt \tag{4.5}$$

where t_0 and T are the start and duration of the analysis frame respectively. Alternative estimates can be found from the first and second *weighted* moments of $f(t)$ using the squared amplitude $[a(t)]^2$ as weight:

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)\, [a(t)]^2\, dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{4.6}$$

$$[B_w]^2 = \frac{\int_{t_0}^{t_0+T} [(\dot{a}(t)/2\pi)^2 + (f(t) - F_w)^2 [a(t)]^2] dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{4.7}$$

where the additional term $(\dot{a}(t)/2\pi)^2$ in B_w accounts for the amplitude modulation contribution to the bandwidth [15].

The following example explains the behavior of F_u vs F_w. Consider a sum $x(t)$ of two sinusoids with constant frequencies $f_1 = 1.5$ kHz and $f_2 = 1.7$ kHz, and time–varying amplitudes $a_1(t)$, $a_2(t)$

$$x(t) = a_1(t) \cos[2\pi f_1 t] + a_2(t) \cos[2\pi f_2 t]\ , \qquad t \in [0, 0.1] \text{ sec} \tag{4.8}$$

where $a_1(t) = 10\,t$, $a_2(t) = 1 - 10\,t$, so that for the first half of the time interval (0 to 50 msecs) the second sinusoid f_2 is dominant, while for the second half (50 to 100 msecs) f_1 dominates. In Fig. 4.1 (a)–(d), we display the amplitude envelope $|a_H(t)|$, $|a_E(t)|$ and the instantaneous frequency $f_H(t)$, $f_E(t)$ of $x(t)$ computed via the HTD and the ESA. The “beating” (in and out of phase) of the two sinusoids manifests itself clearly at the envelope plots (a), (b). At envelope maxima the HTD instantaneous frequency (c) takes the (intuitive) value of $f_H = (a_1 f_1 + a_2 f_2)/(a_1 + a_2)$, while at envelope minima f_H presents spikes of value $f_H = (a_1 f_1 - a_2 f_2)/(a_1 - a_2)$ (see appendix), i.e., the spikes point towards

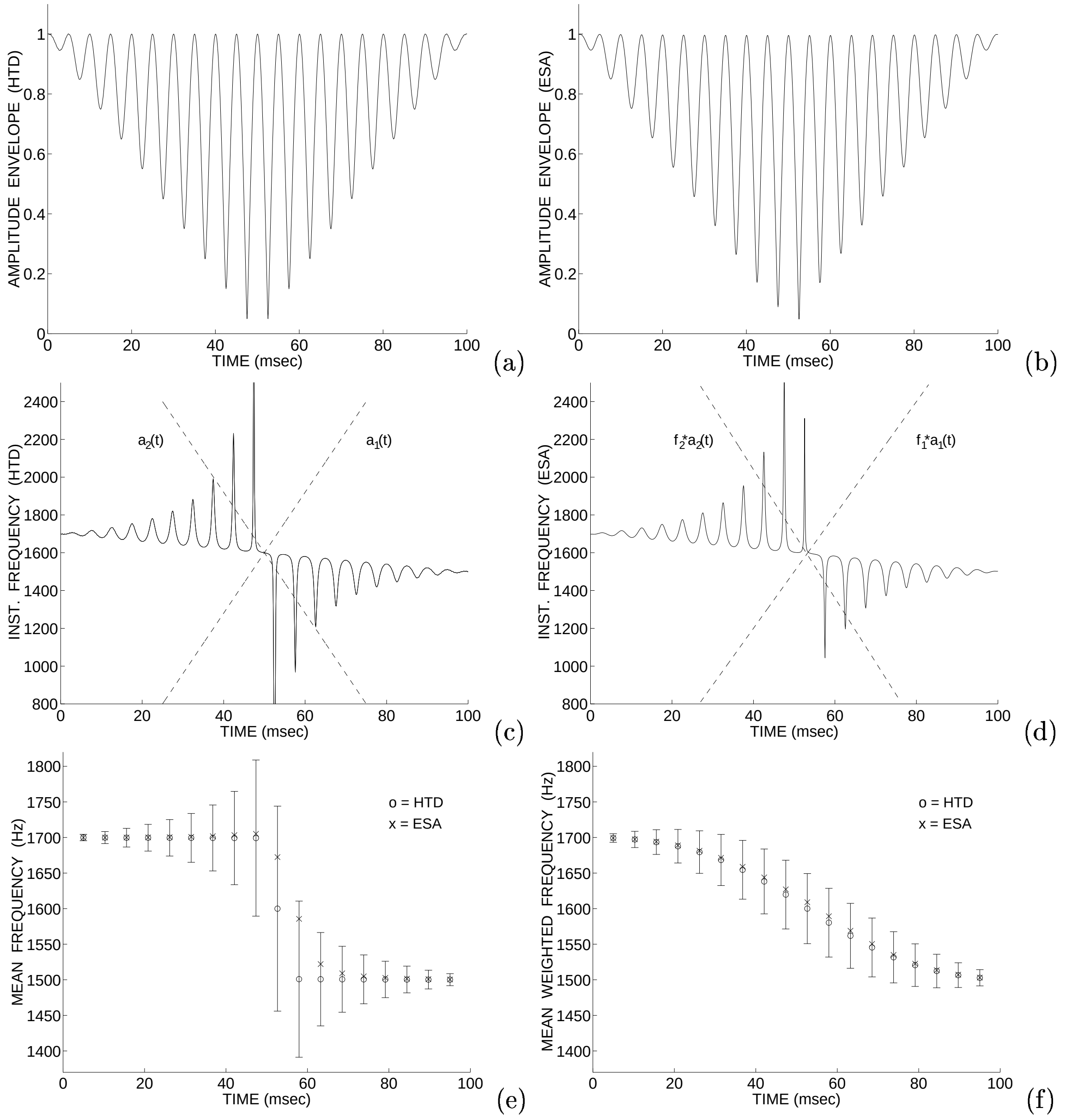


Figure 4.1: Amplitude envelope and instantaneous frequency of $x(t) = a_1(t)\cos[2\pi f_1 t] + a_2(t)\cos[2\pi f_2 t]$, $a_1(t) = 10\,t$, $a_2(t) = 1 - 10\,t$, $t \in [0, 0.1]$ sec, sampled at 10 kHz, estimated via HTD (a), (c) and ESA (b), (d). Dotted lines in (c) are proportional to the amplitude of the sinusoids and in (d) proportional to the amplitude frequency products. Short–time frequency and bandwidth (error bars) estimates: (e) F_u (o is for HTD and × for ESA) and B_u (HTD only), (f) F_w and B_w, for a window of 10 msecs, updated every 5 msecs.

the frequency of the sinusoid with the larger amplitude. The ESA and HTD frequency estimates take similar values, yet the orientation of the instantaneous frequency spikes in (c), (d) is somewhat different. As shown in the appendix, the spikes in the ESA estimate of f_E point toward the frequency of the sinusoid with the larger amplitude frequency product,[2] i.e., the spikes point towards the frequency of the sinusoid produced by the source with the highest energy.

The short–time estimate F_u computed by the ESA and the HTD is shown in Fig. 4.1 (e); F_u locks onto the sinusoid with the greater amplitude (amplitude frequency product for the ESA). The weighted estimate F_w provides a more "natural" short–time formant frequency estimate, because the spikes of the instantaneous frequency correspond to amplitude minima and get weighted less in the F_w average. Actually, F_w is the mean weighted frequency of the two sinusoids, with weight the squared amplitudes. Note that the ESA short–time estimates take slightly greater values than the HTD ones, especially when $a_1 \approx a_2$ (see appendix for explanation).

These results can be generalized to the short–time frequency estimates of speech resonances by use of a sinusoidal speech model. A speech signal can be modeled as a sum of sinusoids with slowly time–varying amplitudes and frequencies [62]; in particular a speech resonance can be modeled as a sum of a few sinusoids. So it is expected that for a speech formant, F_u will have the tendency to lock on the harmonic with the greatest amplitude in the formant band, while F_w will weight the frequency of each harmonic with its squared amplitude (see also appendix).

In Fig. 4.2 (a), we show the Fourier spectrum of a 25 msecs speech segment and the frequency response of the Gabor filter centered around the formant at 1600 Hz with effective RMS bandwidth at 440 Hz. The Fourier spectrum of the formant band signal along with the short time frequency estimates F_u and F_w are shown in (b). Note that F_u locks onto the harmonic with the greatest amplitude in the spectrum, while F_w provides an "average" spectral frequency, a more accurate formant frequency estimate. In Fig. 4.2 (c) and (d), we use a Gabor filter that is centered at 1300 Hz, 300 Hz off the formant frequency. F_u

[2] This can be witnessed in plot (d) where the dashed lines are proportional to the amplitude frequency product for each sinusoid. The turning point for the direction of the instantaneous frequency spikes is where the dashed lines cross.

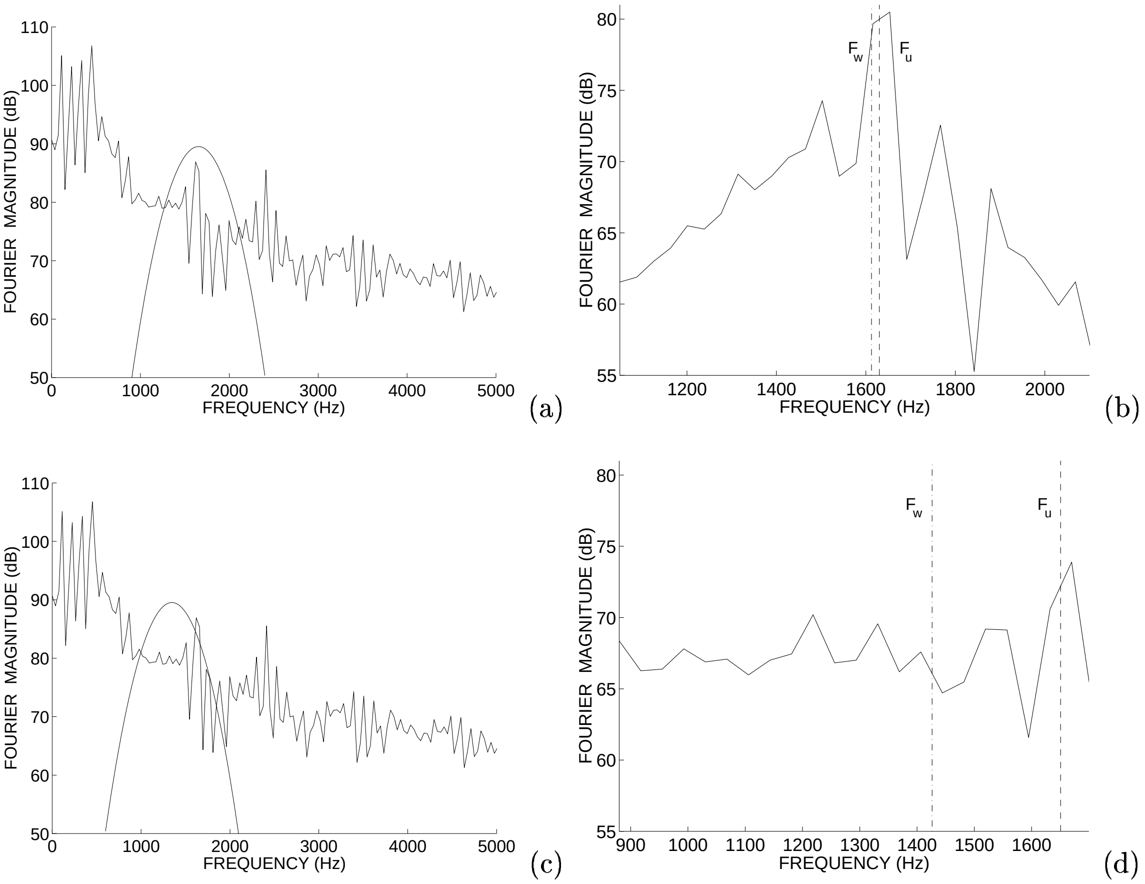


Figure 4.2: (a) The Fourier spectrum of a 25 msecs segment of speech and the frequency response of the Gabor filter centered at $\nu = 1600$ Hz, (b) the Fourier spectrum of the Gabor bandpass filtered speech; the F_u (dashed line) and F_w estimates (dashed–dotted line). (c),(d) same as (a),(b) but now the Gabor filter is centered at $\nu = 1300$ Hz.

still locks on the harmonic with the greatest amplitude in the spectrum, which is in this case the major formant harmonic, while F_w being an "average" frequency deviates from the formant frequency by almost 200 Hz. In this case, the spikes of the instantaneous frequency "lead" towards the formant and the unweighted estimate F_u is a better formant frequency estimate than F_w. There are cases, though, where no major formant harmonics are "inside" the Gabor filter; there the behavior of F_u is unpredictable and thus unstable.

The advantages of the F_u estimate are that it is computationally simple, conceptually attractive, and that it converges faster to the formant frequency in an iterative formant tracking scheme (see for example [26] and Section 4.4). The weighted frequency estimate F_w provides more accurate formant frequencies and is more robust for low energy bands (spectral valleys).

Similarly, the B_w bandwidth estimate is more robust than the B_u estimate. For example, in Fig. 4.1 (e), (f) we display B_u and B_w computed by the HTD for the sum of two sinusoids of Eq. (4.8). The bandwidths are shown as error bars around their respective frequency estimates. Note that for $a_1 \approx a_2$ (i.e., when there is not a single prominent harmonic in the spectrum) B_u takes unnaturally large values. For this reason B_w is used as the formant bandwidth estimate for the remainder of this chapter. Further, B_w is the RMS formant bandwidth.

In [110, 52, 15], the (squared amplitude) weighted estimates F_w and B_w are shown to be time domain equivalents of the first and second central spectral moments of the signal. This explains why the weighted estimates are more robust than the unweighted ones. It also offers an alternative way of computing the F_w and B_w estimates in the frequency domain (see Section 4.6). Note that since B_w equals the second spectral moment, B_w is by definition the RMS bandwidth of the resonance signal.

Overall, the HTD and the ESA provide similar frequency and bandwidth short–time estimates, while, the ESA has smaller computational complexity and better time resolution [84]. As we have noted before, when the center frequency of the Gabor filter approaches the fundamental frequency, the HTD produces smoother estimates than the ESA, when a careful and computationally expensive implementation is used for the discrete–time HTD. In general, the performance of the ESA is expected to deteriorate for frequency bands below

500 Hz because there the carrier frequency (formant) is comparable to the modulation frequency (fundamental) [54]. In practice, the short–time bandwidth estimates B_w for frequency bands in the 0–500 Hz range are more accurate when computed by the HTD. If accurate formant bandwidth estimates are needed in this low frequency range the HTD should be used instead of the ESA for demodulation; otherwise the ESA should be used for computational efficiency.

4.4 Multiband Demodulation Formant Tracking Algorithm

Next, a parallel multiband filtering and demodulation scheme for formant tracking is proposed. The speech signal is filtered through a bank of Gabor bandpass filters, uniformly spaced in frequency with typical effective RMS Gabor filter bandwidth of 400 Hz. The amplitude envelope $|a_E(t)|$ (or $|a_H(t)|$) and the instantaneous frequency $f_E(t)$ (or $f_H(t)$) signals are estimated for each Gabor filter output. Short–time frequency $F_w(t,\nu)$ and bandwidth $B_w(t,\nu)$ estimates are obtained from the instantaneous amplitude and frequency signals (Eqs. (4.6), (4.7)), for each speech frame located around time t and for each Gabor filter centered at frequency ν. The time–frequency distributions $F_w(t,\nu)$, $B_w(t,\nu)$ have time resolution equal to the step of the short–time window (typically 10 msecs) and frequency resolution equal to the center frequency difference of two adjacent filters (typically 50 Hz).

In Fig. 4.3 (c), we plot the value of the short–time frequency estimates $F_w(t,\nu)$ for every frequency band centered at frequency ν vs. time t for the sentence in (a). Note that the y–axis in Fig. 4.3 (c) represents the range of F_w. In (c), the formant tracks are denoted as regions of high plot density (high concentration of frequency estimates), in a similar way that high Fourier amplitudes outline the formant tracks at the speech spectrogram of Fig. 4.3 (b). We refer to the time–frequency representation of Fig. 4.3 (c) as the *speech pyknogram*.[3] The pyknogram displays clearly the formant frequencies and bandwidths, and possibly the location of the spectral zeros (low density areas). Note that a similar time–frequency representation has been proposed in [21], where for each frequency band the instantaneous frequency signal is computed, smoothed in the frequency and time domain and displayed vs. time.

[3] "Pyknogram" stems from the Greek word "pykno" ($\pi\upsilon\kappa\nu\acute{o}\varsigma$) = dense.

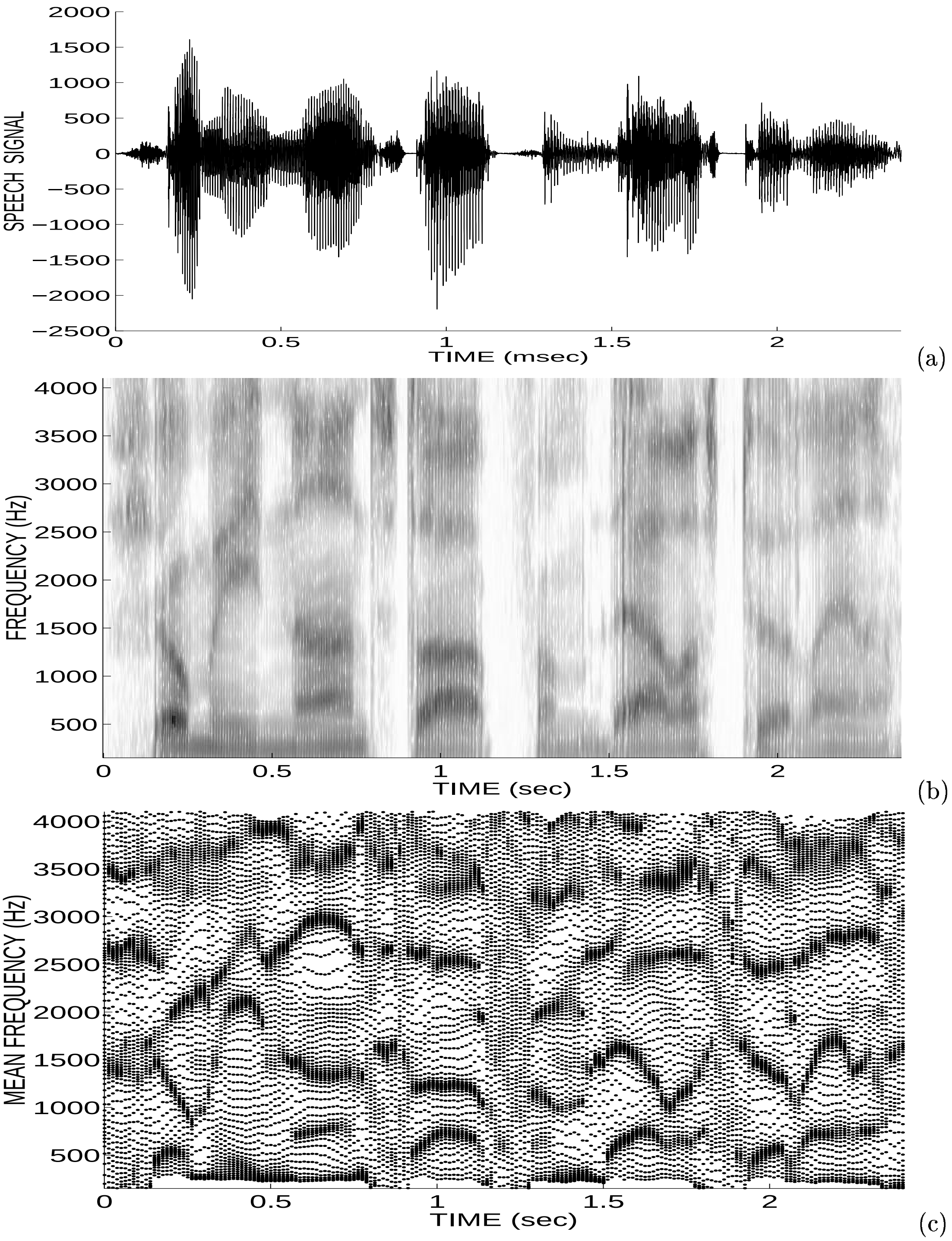


Figure 4.3: (a) Speech signal: "Show me non–stop from Dallas to Atlanta" (b) wideband spectrogram and (c) pyknogram, i.e., the short–time frequency estimates $F_w(t, \nu)$ for the output of 80 Gabor filters spanning $\nu = 200$ to 4200 Hz displayed vs. time (analysis frame update is 12.5 msecs).

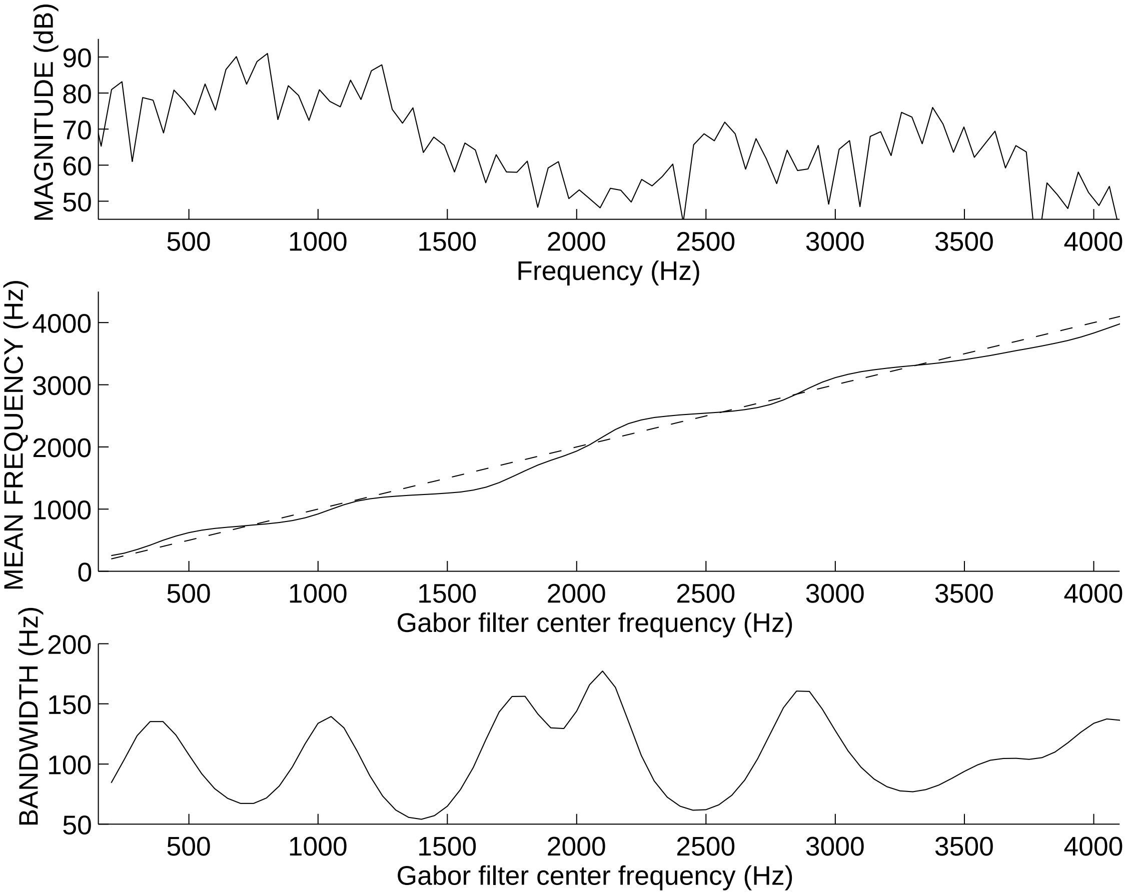


Figure 4.4: The short–time Fourier spectrum, the frequency $F_w(\nu)$ and bandwidth $B_w(\nu)$ estimates vs. the center frequencies ν of the Gabor filters for a 25 msecs frame of speech.

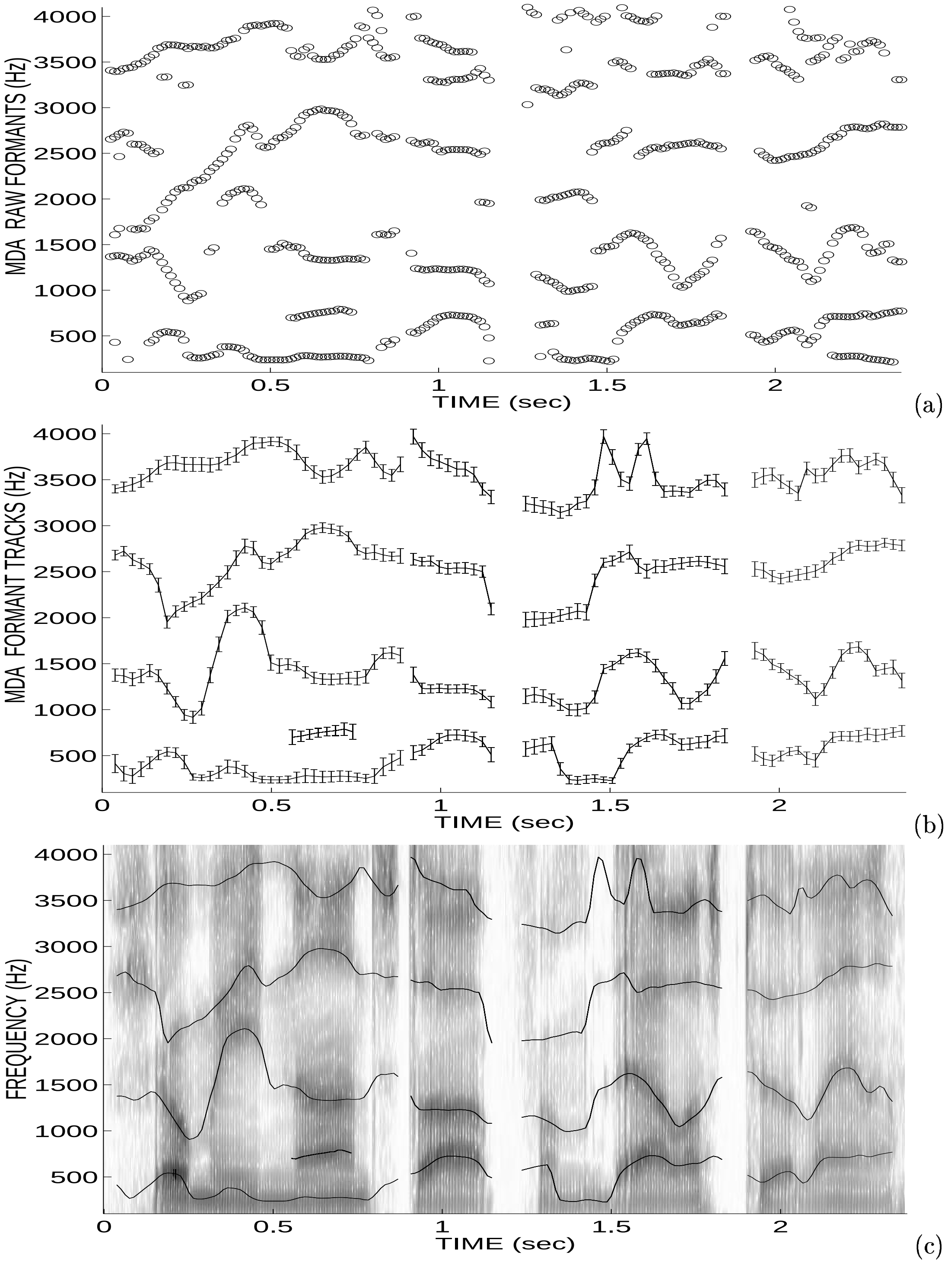


Figure 4.5: MDA formant tracking on real speech: (a) Raw formant estimates, (b) Formant tracks: frequency and bandwidth (error bars) and (c) Formant tracks superimposed on the speech spectrogram.

In Fig. 4.4, we show the frequency $F_w(t, \nu)$ and bandwidth $B_w(t, \nu)$ estimates for a single analysis frame centered at t vs. the center frequency of the Gabor filters ν. Note that the speech resonances in the Fourier spectrum approximately correspond to points where the Gabor filter center frequency ν and the short time frequency estimate $F_w(\nu)$ are equal, i.e., $F_w(\nu) = \nu$. These are points where the solid line (frequency estimate) meets the dotted one (Gabor filter center frequency). In addition, we have observed that bandwidth $B_w(\nu)$ minima also indicate the presence of formants.

A simple way to define raw formant estimates is as the frequency where the Gabor filter center frequency ν and the short time frequency estimate $F_w(\nu)$ are equal, i.e., $F_w(\nu) = \nu$. Yet, we have observed from synthetic and real speech experiments that for a "weak" formant the $\{\nu : F_w(\nu) = \nu\}$ frequency estimate is biased towards the frequency of a neighboring "strong" formant. As a result, the second and higher formant estimates may be inaccurate, especially, when the separation of two formant tracks is small. More accurate formant estimates are obtained from the value of $F_w(\nu)$ at inflection points where $\partial^2 F_w(\nu)/\partial\nu^2 = 0$. Inflection points of $F_w(\nu)$ correspond to dense regions of the pyknogram, since the slope $\partial F_w(\nu)/\partial\nu|_{\nu_0}$, that is a measure of the concentration of frequency estimates around ν_0, has minima there. For best results a hybrid raw formant decision is used: the more robust $\{\nu : F_w(\nu) = \nu\}$ estimate for $\nu < 500$ Hz and $\{F_w(\nu) : \partial^2 F_w(\nu)/\partial\nu^2 = 0\}$ for $\nu > 500$ Hz.

For the raw formant at $F_w(\nu_0)$ the slope of $F_w(\nu)$ at ν_0 determines the prominence of the formant candidate. As the slope $\partial F_w(\nu)/\partial\nu|_{\nu_0}$ approaches zero, the short–time frequency estimate $F_w(\nu)$ becomes almost constant for bands located around ν_0, a sign that a "strong" formant peak exists in the vicinity. Clearly the slope for a legitimate formant candidate ranges from zero (most probable candidate) to one (least probable candidate). One may either use $\partial F_w(\nu)/\partial\nu$ as a weight in the formant tracking decision algorithm or a threshold (typically 0.6 to 0.8) can be imposed on the slope. We have implemented the latter approach with good results, i.e., only formant candidates with slopes below 0.7 are selected as raw formants. The former approach, although more complicated, is attractive and should be investigated in the future.

In brief, for each speech analysis frame centered at time t the raw formants RF are obtained from the time–frequency distribution $F_w(t,\nu)$ as follows:

$$\begin{aligned} RF_1 &= \{\nu : \ (F_w(\nu) = \nu) \ \text{and} \ (\frac{\partial F_w(\nu)}{\partial \nu} < 0.7) \ \text{and} \ (\nu < 500)\} && (4.9) \\ RF_2 &= \{F_w(\nu) : \ (\frac{\partial^2 F_w(\nu)}{\partial \nu^2} = 0) \ \text{and} \ (\frac{\partial F_w(\nu)}{\partial \nu} < 0.7) \ \text{and} \ (\nu > 500)\} && (4.10) \end{aligned}$$

$$RF = RF_1 \bigcup RF_2 \qquad (4.11)$$

where $\bigcup$ denotes set union.

In Fig. 4.5 (a), we display the raw formant estimates for the sentence of Fig. 4.3 (a). A 3–point binomial smoother is applied on $F_w(t,\nu)$ in the time domain before the raw formant estimates are computed. In Fig. 4.5 (b), (c), the formant tracks (frequency and bandwidth) are shown superimposed on the speech spectrogram. Formant bandwidths are obtained from the B_w estimate. Note that B_w is an estimate of the RMS formant bandwidth.

The decision algorithm used to convert raw formants to formant tracks is similar to linear prediction (LP) based formant tracking algorithms [65]. Special care is taken for nasals sounds where a "nasal formant" between the first and second formant is allowed to be "born and to die". First, we search for anchor formant segments, i.e., segments where the formant tracks are well separated in frequency and well defined. Next, the formant tracks between anchor segments are filled using continuity constraints. Finally, we determine if a "nasal formant" is present between the first and the second formant tracks. The decision algorithm is kept simple since the number of spurious raw formants is very small. In general, the choice of a decision algorithm depends on the application. In our case, the formant tracks are used for vocoding so the decision algorithm is tuned to guarantee continuous formant tracks [83]. Alternative approaches such as an exhaustive search of all possible combination of raw formants to formant tracks (e.g., Viterbi decoding [47]) or a functional minimization approach [50] can be found in the literature.

4.5 Performance and Comparisons

The multiband demodulation analysis (MDA) formant tracking algorithm was tested on synthetic speech signals produced by a cascade formant synthesizer. An example is displayed in Fig. 4.6. Speech was synthesized using the tracks shown as dotted lines in Fig. 4.6 (b).

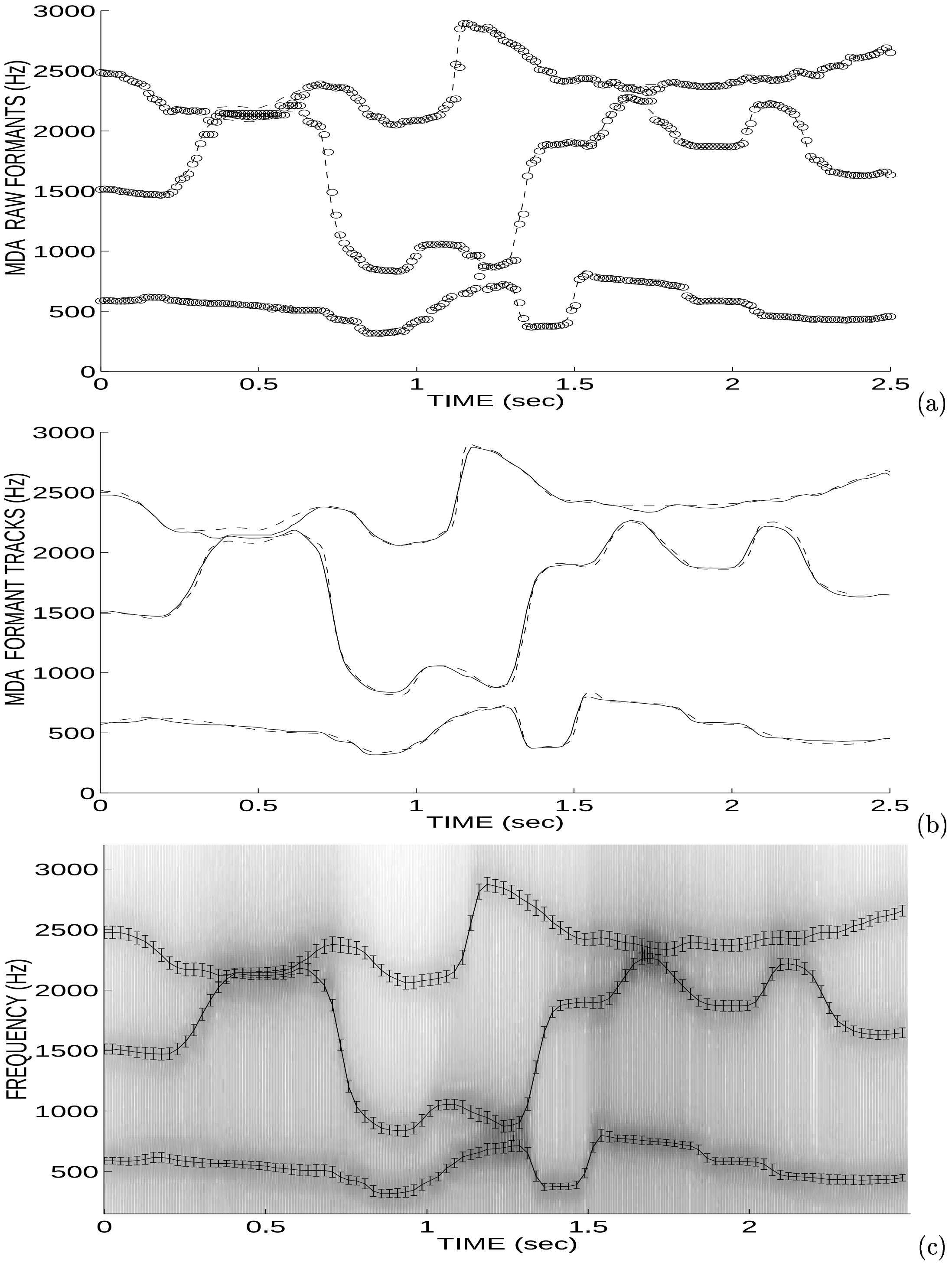


Figure 4.6: MDA formant tracking on synthetic speech: (a) Raw formant estimates, (b) Formant tracks: computed (solid line) vs. actual (dotted line) (c) Formant tracks superimposed on the speech spectrogram.

The formant trajectories were designed by hand (nonsense utterance) and their 3 dB bandwidths were constant throughout the synthetic utterance at 60, 70 and 80 Hz for the three formants. The MDA raw formant estimates are shown in Fig. 4.6 (a) and the resulting formant tracks are shown at (b), (c) as solid lines. The algorithm produced good formant estimates and was able to accurately track rapidly evolving formant tracks and weak formants. Formant merging occurred for frequency separation less than approximately 150 Hz, as shown for the second and third tracks in Fig. 4.6 (b). In this case, increased frequency discrimination can be obtained by decreasing the bandwidths of the filters in the filterbank. The formant bandwidth estimates shown as error bars in Fig. 4.6 (c) were also accurate. An empirically determined bandwidth correction factor was applied in regions where formant variations were greater than 100 Hz/10 msecs to compensate for overestimated bandwidth values.

Overall, the MDA produced accurate formant frequency and bandwidth estimates for synthetic speech. The formant estimates were more accurate for lower than for higher fundamental frequency values. In general, when the fundamental frequency is comparable to the bandwidth of the Gabor filter, only a single speech harmonic “falls inside” the filter and the MDA tracks the most prominent harmonic in the formant band instead of the formant frequency. In this case, the bandwidth estimates are also noisy. For high–pitched synthetic speech more accurate formant tracks can be obtained by increasing the bandwidth of the Gabor filters. In general, when choosing the filter bandwidth the tradeoff between increased frequency discrimination and accurate formant estimates for high–pitched speakers should be considered carefully.

Next, the formant tracking algorithm was tested on clean and on telephone speech from the TIMIT and NTIMIT databases, respectively, with good results. The quality of the formant tracks was determined by superimposing the estimated formant trajectories on the speech spectrogram. The formant frequency and bandwidth estimates were accurate in all cases except for high–pitched female speakers. Further, the performance of the algorithm on telephone speech sentences (NTIMIT) was good. The estimated formant tracks were similar to the ones obtained from the corresponding high–quality TIMIT sentences. Problems occurred for the third formant track when it exceeded 2500 Hz due to the bandpass filtering effects of the telephone channel. Also, weak formant tracks were sometimes inaccurate or lost due to noise. Overall, the MDA formant tracking performed well for both clean and telephone speech.

Most formant tracking algorithms are based on a short–time linear prediction (LP) analysis. LP is a parametric method that computes a predetermined number of formant estimates independent of the actual number of spectral peaks in the spectrum. In addition, the formant frequency accuracy is affected by the preemphasis, while the harmonic structure of the spectrum and the formant bandwidth estimates are unrealistic. Finally, LP–based formant trackers encounter problems with nasals and nasalized vowels. The multiband demodulation approach overcomes most of these problems. In Fig. 4.7, we display the LP raw formant frequency and bandwidth estimates for comparison with the MDA estimates in Fig. 4.5. Although the long–term formant trajectory shapes look similar (except for nasalized speech, where the MDA sometimes produces an extra low formant[4]), there are some important differences over small scales. LP produces a number of spurious formants that make the formant decision algorithm more complex. Also, the LP raw formants estimates are noisy, especially for weak and/or higher formants. Finally, in (b) the LP bandwidth estimates (shown as error bars, scaled up four times) are inaccurate and very noisy. Overall, the MDA formant tracking algorithm has the attractive features of being conceptually simple and easy to implement in parallel. It behaves well in the presence of nasalization, provides realistic formant bandwidth estimates and produces very few spurious raw formants.

An iterative demodulation algorithm for formant tracking has been proposed by Hanson, Maragos and Potamianos in [26]. Initial formant estimates are refined through an iterative scheme: a Gabor bandpass filter is centered at the initial formant estimate; the speech resonance is extracted through filtering, demodulated, and the short–time mean frequency F_u is computed. At the next iteration the Gabor filter center frequency is set to the formant estimate F_u. The algorithm converges to a formant when F_u does not change significantly from one iteration to the next. For the iterative ESA the F_u frequency estimate is preferred over F_w because the use of F_u increases substantially the convergence speed to a formant. Overall, the MDA produces better formant estimates that the iterative ESA especially in

[4] This extra formant can be either eliminated by the decision algorithm or preserved as an extra formant track, e.g., in a vocoding application (see Chapter 6).

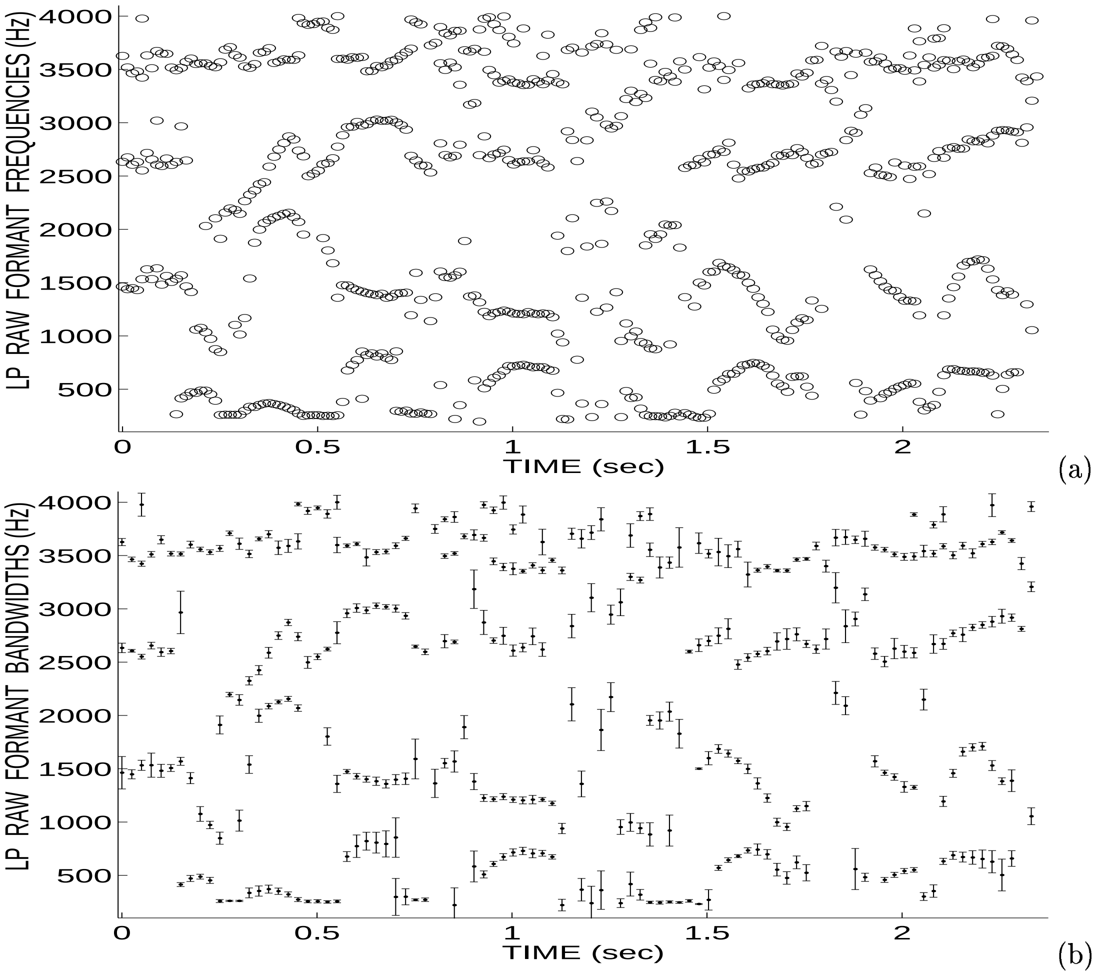


Figure 4.7: LP raw formant frequency (a) and bandwidth (scaled up 4 times) (b) estimates for the speech signal shown in Fig. 3(a); LP analysis order is 12, preemphasis is 0.5, window size is 25 msecs updated every 12.5 msecs.

regions when the separation between formant tracks is small. This is due to the improved raw formant decision algorithm of the MDA. A modified iterative ESA algorithm that uses gradient descent to reach the local minima of $\partial F_u(\nu)/\partial\nu$ could significantly improve the accuracy of the formant tracks produced by the iterative ESA.

4.6 Discussion

The multiband demodulation formant tracking algorithm uses a bank of uniformly spaced Gabor filters. Alternatively, for a small additional computational cost, a Gabor wavelet (constant–Q filterbank) can be used, which gives uniform performance for ESA demodulation across channels [9]. Increasing the spacing of the bandpass filters with frequency, decreases the frequency resolution for higher formants estimates. This is compatible with the perceptual resolution (limens) of the ear. In [26], the performance of the iterative ESA formant tracker has improved by using constant–Q filters.[5]

As discussed in Section 4.3, the choice of the unweighted F_u or the weighted F_w frequency estimate is mainly the choice between fast "convergence" to a formant and robust raw formant estimates. In general, for the MDA formant tracking algorithm we prefer to use the more reliable weighted estimate F_w. When the frequency axis is poorly sampled (i.e., when only a few Gabor filters are used), though, F_u can produce better results than F_w, since F_u provides good formant estimates even when the Gabor filter is not centered exactly on the formant frequency.

We mentioned in Section 4.3 that the estimates F_w and B_w can be computed in the frequency domain as the first and second spectral moments, e.g., using the fast Fourier transform (FFT). This results in significant computational savings since the Gabor filtering can be implemented by multiplication in the frequency domain and no demodulation is needed. The F_w and B_w estimates computed in the frequency domain take similar values to their time domain equivalents when adequately "long" FFT implementation is used. A 1024–point FFT gives good results for sampling frequency at 16 kHz and a short–time

[5] In the next chapter a Mel–spaced filterbank is used for multiband demodulation pitch estimation. The constant–Q Mel–spaced filterbank is shown to outperform the uniformly spaced one for the pitch tracking application (see Section 5.5).

analysis window of 20 msecs. From our simulations on synthetic speech, though, we have observed that the time domain implementation is able to better resolve “weak” formant regions. In addition, when using the time domain implementation, one may enhance the formant track time resolution at a small computation cost by simply decreasing the size of the short–time averaging window in a second pass of the algorithm.

An alternative raw formant decision algorithm is to use image processing techniques directly on the speech pyknogram. The information in the pyknogram can be mathematically represented as a two dimensional set in the time–frequency plane. As seen from Fig. 4.3 (c), the formant tracks manifest themselves as relatively thin and elongated geometrical structures. Formant tracking can be performed on the pyknogram by cleaning these dense regions from the surrounding clutter and thining them down to a single point at each time instant. Such a geometrical analysis of the pyknogram can be rigorously quantified using the concepts and operations of mathematical morphology. This is a powerful set–theoretic methodology for image analysis that can quantify the shape, size, and other geometrical aspects of image objects, which has found many applications in image processing and nonlinear filtering [98, 60]. As a continuation of the work in this chapter, we plan to apply algorithms from morphological image analysis for cleaning, segmentation, and thinning of the formant tracks in the pyknogram.

Finally, one may possibly use multiband demodulation for spectral zero tracking. In the speech pyknogram of Fig. 4.3 (c), zeros sometimes manifest themselves as areas of low plot density. For example, for nasalized sounds an anti–resonance can be observed between the second and the third formant track in (c). More work is underway for anti–resonance tracking using multiband demodulation.

4.7 Conclusions

In this chapter, we have presented a collection of ideas and algorithms for estimating speech formants and tracking their evolution in time based on the AM–FM speech model and demodulation algorithms. The main speech analysis tool used was multiband filtering followed by demodulation (MDA). We have seen that the proposed MDA formant tracking algorithm produces good formant frequency and bandwidth estimates for synthetic, clean and telephone speech, while overcoming most of the drawbacks of LP–based formant trackers. In addition, we demonstrated that the MDA approach is a powerful speech analysis tool that produces rich time–frequency representations such as the speech pyknogram. Further, in this paper, we have compared the unweighted mean and the (squared amplitude) weighted mean of the instantaneous frequency for formant frequency estimation. We concluded that the weighted estimate provides in general more reliable and accurate formant locations. The unweighted mean is preferred when the filter (used for extracting the formant from the spectrum) is positioned far from the formant or for increased convergence speed in an iterative formant tracking scheme.

Overall, the multiband demodulation formant tracker produced very promising results, which suggests that the AM–FM modulation model and the energy demodulation algorithms are a useful modeling approach for speech analysis.

Appendix

Consider the sum of two or more sinusoids with time–varying amplitudes $a_n(t)$ and constant frequencies f_n[6]

$$x(t) = \sum_n a_n(t) \, \cos[2\pi f_n t + \theta_n] \tag{4.12}$$

where θ_n are arbitrary phase constants. Assuming that the bandwidth of $x(t)$ is much smaller than $\min_n(f_n)$, the quadrature error (Eq. (3.8)) will be small and the Gabor analytic signal $z(t)$ of $x(t)$ will be

$$z(t) \approx \sum_n a_n(t) \, \exp[j\,(2\pi f_n t + \theta_n)]. \tag{4.13}$$

Under the additional assumption that $a_n(t)$ is slowly varying compared to $\cos[2\pi f_n t]$, the HTD estimates for the amplitude envelope $|a_H(t)|$ and instantaneous frequency $f_H(t)$ are

$$|a_H(t)| = |z(t)| \approx (\, \textstyle\sum_{n,k} \, a_n(t)\, a_k(t)\, \cos[2\pi(f_n - f_k)t + (\theta_n - \theta_k)]\,)^{\frac{1}{2}} \tag{4.14}$$

$$f_H(t) = \tfrac{d}{dt}\angle z(t) \approx \textstyle\sum_{n,k} \, f_n \, a_n(t)\, a_k(t)\, \cos[2\pi(f_n - f_k)t + (\theta_n - \theta_k)]/[a_H(t)]^2 \tag{4.15}$$

For the case of two sinusoids (we set $\theta_1 = \theta_2 = 0$ for simplicity)

$$|a_H(t)| = (\, a_1^2 + a_2^2 + 2\, a_1\, a_2 \, \cos[\Delta\omega\, t]\,)^{\frac{1}{2}} \tag{4.16}$$

$$f_H(t) = (\, a_1^2\, f_1 + a_2^2\, f_2 + a_1\, a_2\, (f_1 + f_2)\, \cos[\Delta\omega\, t]\,) \,/\, a_H^2 \tag{4.17}$$

where $\Delta\omega = 2\pi(f_1 - f_2)$. At envelope maxima and minima ($\cos[\Delta\omega\, t] = \pm 1$), $|a_H|$ and f_H take the values

$$|a_H| = |a_1 \pm a_2| \qquad\qquad f_H = \frac{a_1\, f_1 \pm a_2\, f_2}{a_1 \pm a_2}. \tag{4.18}$$

Thus, at envelope minima f_H presents spikes pointing towards the frequency of the sinusoid with the larger amplitude a_n. From Eqs. (4.16) and (4.17) the short time frequency estimates F_u and F_w defined in Eqs. (4.6) and (4.7) respectively, are (after some algebra proven to be) approximately equal to

$$F_u \approx \begin{cases} f_1, & a_1 > a_2 \\ f_2, & a_1 < a_2 \end{cases} \qquad\qquad F_w \approx \frac{a_1^2\, f_1 + a_2^2\, f_2}{a_1^2 + a_2^2} \tag{4.19}$$

[6] The analysis that follows also holds for an additional slow–varying phase modulation term, i.e., for a sum of amplitude and frequency modulated sinusoids.

i.e., F_u locks onto the frequency component with the larger amplitude, while F_w provides a (squared amplitude) weighted mean frequency. The exact values of F_u and F_w depend on the analysis frame boundaries.

One can obtain equations similar to (4.14), (4.15) for the ESA but they are of little intuitive value. Instead we concentrate on the sum of two amplitude modulated sinusoids. The value of the amplitude envelope $|a_E|$ and instantaneous frequency f_E at envelope maxima and minima derived from the continuous time ESA are

$$|a_E| = |a_1 \pm a_2| \qquad\qquad f_E = \left| \frac{a_1 f_1^2 \pm a_2 f_2^2}{a_1 \pm a_2} \right|^{\frac{1}{2}} . \tag{4.20}$$

As a result, the frequency presents spikes at envelope minima that point toward the frequency of the sinusoid with the larger amplitude frequency product, i.e., $a_n f_n$. Similarly, the short time estimate F_u is approximately equal to the frequency of the sinusoid with the larger amplitude frequency product $a_n f_n$. Finally, F_w takes values similar to Eq. (4.19).

The F_w estimate computed using the ESA takes slightly higher values than F_w computed using the HTD, especially for $a_1 \approx a_2$. This is due to the increased frequency weighting in Eq. (4.20) compared to Eq. (4.18). When $|f_1 - f_2| \ll f_1, f_2$ the differences between ESA and HTD can be ignored for all practical purposes. Similarly, the performance of the ESA and the HTD are almost identical for speech formant demodulation when the fundamental frequency is much smaller than the formant frequency. For demodulation of a large bandwidth multi–component signal, though, the two algorithms can produce quite different results. There, the ESA frequency estimates are biased [54] (the ESA overestimates the frequencies as can be seen from comparing Eqs. (4.18) and (4.20)).

For a sum of more than two (AM–FM) sinusoids: $F_w \approx (\sum_n a_n^2 f_n) / (\sum_n a_n^2)$ (directly from Eqs. (4.14), (4.15)), i.e., F_w weights each frequency with the squared amplitude. The behavior of F_u is more complicated. In general, if the signal contains only one or two prominent sinusoids, F_u will lock onto the frequency of the sinusoid with the greatest amplitude. This is typically the case for a speech resonance signal.

Chapter 5

Speech Analysis II: Fundamental Frequency Estimation

5.1 Introduction

Estimating the fundamental frequency of voiced speech is a challenging problem because speech is not a perfectly periodic signal. Further, the excitation is often "corrupted" with secondary pulses and/or aspiration noise that destabilize the pitch estimation procedure. Finally, it is hard to obtain robust fundamental frequency estimates for unstable voiced speech segments, e.g., phonemic transitions, where the periodicity of the excitation is not well established. Most pitch tracking algorithms estimate an average pitch that is defined over a speech segment of 2–3 pitch periods. The average pitch estimate contains crucial information about the source of voicing, although, it often fails to fully characterize the excitation signal. Overall, fundamental frequency estimation is an important speech analysis application partly because most low bit rate vocoders use a short–time pitch estimation/prediction scheme to effectively code and reconstruct the speech waveform [62, 96].

The fundamental frequency can be estimated either in the time domain or in the frequency domain. Most time domain algorithms attempt to measure the similarity between consecutive pitch periods, e.g., from the maxima of the autocorrelation function [92] or the average magnitude difference function (AMDF) [94]. For increased robustness, the adverse effects of vocal tract filtering and additive noise on the speech signal can be modeled and accounted for before the pitch is estimated. The SIFT algorithm [61] subtracts some of the vocal tract contribution to the speech signal by inverse linear prediction (LP) filtering and then seeks the maxima of the autocorrelation function of the error signal.[1] Further, for increased robustness to additive Gaussian noise the maximum likelihood pitch estimator can be used [111]. An alternative view of the time–domain fundamental frequency estimation/prediction techniques is considered in [71, 48]. A simple model is proposed to account for quasi–periodicity in speech, e.g., $s(t) = a\ s(t-\tau)$, where τ is the pitch period; pitch is computed by minimizing the modeling error function.

Frequency domain pitch estimation algorithms explore the harmonic structure of the short–time Fourier transform of voiced speech. A simple and efficient pitch estimation procedure is peak picking of the speech cepstrum[2] [74]. Another interesting pitch estimation scheme was proposed in [63, 23] in the context of the sinusoidal model [62]. First, the frequency of each harmonic in the short–time Fourier spectrum is computed. Next, for each fundamental frequency candidate F_0, a functional measures the deviation of multiples of the fundamental frequency candidates kF_0 from the estimated kth harmonic frequency. Pitch is estimated by minimization of this functional over all possible candidates F_0.[3]

In this chapter, multiband demodulation analysis (MDA) is applied to the problem of fundamental frequency estimation. In Section 5.2, we show that under certain conditions the short–time average of the instantaneous frequency F_u (Eq. (4.4)) is an accurate estimate of the most prominent spectral component in the speech band. Specifically, for voiced speech F_u locks on a harmonic, i.e., a multiple of the fundamental frequency. Based on this observation we introduce the *multiband demodulation pitch estimation algorithm* in Section 5.3. Multiband demodulation (see Section 4.2) is used to compute the instantaneous frequency signal $f(t)$ and a short–time harmonic frequency estimate for each frequency band. The estimated harmonic frequencies are incorporated in a functional minimization procedure that determines the pitch. Implementation details, computational issues, and relations with other algorithms are discussed in Sections 5.4 and 5.5. Overall, the multiband

[1] The LP error signal is a rough estimate of the speech excitation signal.

[2] The cepstrum is defined as the inverse transform of the logarithm of the Fourier transform of the signal.

[3] Notably, a "parallel" pitch period estimator was first proposed in [24]; there pitch was computed from two different frequency bands for increased robustness.

demodulation pitch tracker is conceptually simple, produces very smooth and accurate pitch contours, and combines the advantages of both time- and frequency–domain algorithms.

5.2 Harmonic Frequency Estimates

In Section 3.4, we have studied the time evolution of the amplitude envelope $a(t)$ and instantaneous frequency $f(t)$ signals for resonances of synthetic and real speech. Here, we elaborate on the detailed shape of the phase signal $\phi(t)$ of a synthetic voiced speech resonance. The end goal is to obtain robust and accurate short–time estimates of the harmonic frequencies kF_0, $k \in \mathcal{N}$, where F_0 is the fundamental frequency. Note that the phase is obtained simply from the integral of the instantaneous frequency, i.e., $\phi(t) = 2\pi \int_{-\infty}^{t} f(\tau)\, d\tau$.

Next, the behavior of the phase signal $\phi(t)$ of a speech resonance is outlined with an example. A synthetic resonance signal with formant frequency at $F = 300$ Hz is produced from a linear second–order system. The system is excited with a periodic train of unit pulses with a fundamental frequency of $F_0 = 200$ Hz.[4] The excitation and the synthetic resonance signals are shown in Fig. 5.1 (a) and (b), respectively. Next, the instantaneous frequency $f(t)$ and phase $\phi(t)$ signals of the resonance are estimated using the ESA; $f_E(t)$, $\phi_E(t)$ are displayed in Fig. 5.1 (c) and (d), respectively. The instantaneous frequency signal $f_E(t)$ is equal to the formant frequency F everywhere, apart from the excitation instants where it presents a single spike. Similarly, the phase $\phi_E(t)$ has a "local" slope equal to F and presents step discontinuities at excitation instants. Assuming that the resonance signal is perfectly periodic[5] the phase modulo 2π is constant at the beginning of each pitch period, i.e., $\phi(n/F_0) \bmod 2\pi = ct$, where $n \in \mathcal{N}$ and ct is an arbitrary constant. As a result, the phase slope in a window containing a few pitch periods is approximately equal to a multiple of the fundamental frequency, depending on the boundaries of the analysis window. We refer to the phase slope computed over a few pitch periods as the "global" or "macroscopic"

[4] The values of F, F_0 are chosen such for better visualization purposes.

[5] This assumption holds for this ideal synthetic speech example but it is only an approximation for real speech resonances.

slope. For example, for an analysis window that is l pitch periods long, beginning at n/F_0

$$(\phi((n+l)/F_0) - \phi(n/F_0)) \bmod 2\pi = 0 \implies \frac{1}{2\pi} \frac{\phi((n+l)/F_0) - \phi(n/F_0)}{(n+l)/F_0 - n/F_0} = kF_0 \quad (5.1)$$

where $k \in \mathcal{N}$. In our case, $k = 2$ because the second harmonic $2F_0 = 400$ Hz is the most prominent harmonic in the formant band. In general, for analysis windows that are not pitch period multiples Eq. (5.1) holds only approximately. The local slope, equal to the formant frequency, and the global slope, equal to the second harmonic, are shown in Fig. 5.1 (d) with dashed–dotted and dashed lines respectively.

In the previous chapter, we were interested in the local slope of the phase, i.e., the formant frequency, and for that reason we used the weighted mean instantaneous frequency F_w (Eq. (4.6)) that deemphasizes the phase discontinuities. Here, we are interested in tracking the speech harmonics and use the unweighted mean instantaneous frequency estimate F_u. The mean instantaneous frequency F_u and the global phase slope as defined in Eq. (5.1) are equal because

$$F_u = \frac{1}{T} \int_{t_0}^{t_0+T} f(t)\, dt = \frac{1}{2\pi} \frac{\phi(t_0+T) - \phi(t_0)}{(t_0+T) - t_0} \quad (5.2)$$

where t_0 and T are the start and duration of the speech segment. As discussed in Section 4.3 and in the previous paragraph, F_u has the tendency to lock onto the most prominent harmonic in the spectrum. In Fig. 4.1 and Fig. 4.2, we have displayed this behavior both for a sum of two amplitude modulated sinusoids and for a speech resonance signal. We have also observed, though, that when two or more spectral components of comparable amplitude exist in the Fourier spectrum of the resonance signal, the value of F_u is undetermined and lies in between the frequencies of the two spectral components. To remedy this problem, we modify the multiband demodulation scheme. Specifically, we choose the spacing and bandwidths of the Gabor filters such that, in the great majority of cases, only a single prominent spectral component exists in each band. Effective RMS Gabor bandwidths around 200 Hz produce good results for the fundamental frequency estimation scheme described in the next section.

Alternatively, one may compute the phase slope by using a first order polynomial fit on the phase signal $\phi(t)$. There is a simple relation between the slope of the phase S_ϕ

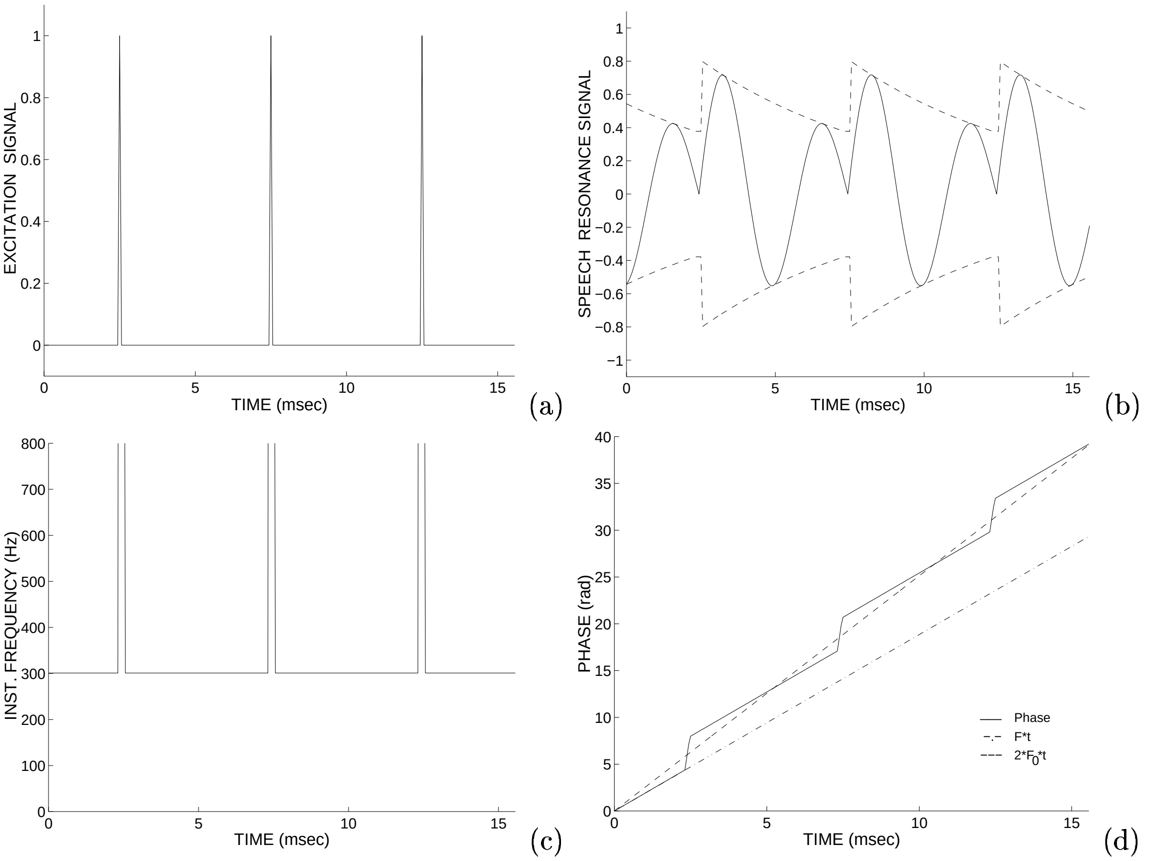


Figure 5.1: (a) Pulse train excitation signal with fundamental frequency $F_0 = 200$ Hz and (b) synthetic speech resonance signal with formant frequency $F = 300$ Hz (amplitude envelope shown with a dashed line). Instantaneous frequency (c) and phase (d) signals for this resonance; the dashed line has slope equal to twice the fundamental frequency $2F_0$, while the slope of the dashed–dotted line is the formant frequency F.

computed from linear regression and the instantaneous frequency $f(t)$

$$S_\phi = \frac{1}{2\pi} \frac{\int_{t_0}^{t_0+T} \; t\,\phi(t)\;dt}{\int_{t_0}^{t_0+T} \;\; t^2\;dt} = \frac{\int_{t_0}^{t_0+T} \; t\;\left[\int_{-\infty}^{t} f(\tau)d\tau\right]\;dt}{\int_{t_0}^{t_0+T} \;\; t^2\;dt} \tag{5.3}$$

where t_0 and T are the start and duration of the analysis window respectively. In general, S_ϕ and F_u take similar values as should be expected from Eqs. (5.3), (5.1), yet, S_ϕ is a more accurate harmonic frequency estimate because it is less sensitive to the choice of the analysis window boundaries and to noisy speech. This is due to the fact that S_ϕ is the slope estimate that minimizes the mean square error, while F_u is computed from the value of the phase at the two boundary points of the analysis window. Henceforth, we adopt S_ϕ as the harmonic frequency estimate.

A final issue is the choice of the demodulation algorithm used to estimate the phase slope S_ϕ. We have discussed in Section 4.3 that the HTD can produce smoother $f(t)$ and F_u estimates than the ESA for bands in the 0–500 Hz frequency range. Similarly, in this low frequency range the S_ϕ estimate is more accurate when computed by the HTD than by the ESA.[6] The problems that the ESA encounters for low frequency bands are greatly alleviated for narrowband signals. For a carefully designed filterbank, e.g., Mel–spaced filters [91] with average bandwidth of 200 Hz, the HTD and ESA phase slope estimates take similar values. The ESA is preferred for demodulation in this chapter because it is computationally simpler than the HTD.

5.3 Multiband Demodulation Pitch Tracking

Next, we propose a parallel multiband demodulation scheme for fundamental frequency estimation. The speech signal is analyzed through a bank of Mel–spaced constant–Q Gabor bandpass filters, with average distance between consecutive filters of 100 Hz, and average effective RMS bandwidth of 200 Hz. The amplitude envelope $a(t)$ and phase $\phi(t) = 2\pi \int_{-\infty}^{t} f(\tau)\,d\tau$ signals are estimated for each band, and the short–time phase slope S_ϕ is computed using a first order polynomial fit (see Eq. (5.3)). This results in a phase slope time–frequency distribution $S_\phi(t, \nu)$ computed at each speech analysis frame located

[6] The ESA sometimes estimates inaccurately the values of the instantaneous frequency spikes or equivalently the phase step discontinuities.

around time t and at each Gabor filter center frequency ν. Typical duration of the short–time analysis frame is 15–20 msecs, with a 10 msecs update.

In Fig. 5.2 (a), we display the Mel–spaced Gabor filterbank used for the multiband filtering stage; 20 filters covering the range 100–2000 Hz are used. The phase slope estimates for each of the 20 Gabor bands are shown in (b) as dashed lines superimposed on the Fourier spectrum for a 20 msecs speech segment of the phoneme /aa/. Note that S_ϕ locks onto the speech harmonics.

In Fig. 5.3 (c), we display the phase slope $S_\phi(t,\nu)$ time–frequency distribution for the TIMIT sentence "Cats and dogs each hate the other" shown in (a). The values of S_ϕ for all 20 bands (see Fig. 5.2 (a)) are plotted vs. time. Note that the prominent speech harmonics are outlined in Fig. 5.3 (c). When compared to the narrowband speech spectrogram (shown in (b)), the harmonics in the phase slope $S_\phi(t,\nu)$ distribution are thinned and more clearly defined.

The fundamental frequency of a voiced speech segment is determined through an exhaustive search of all the possible candidates in the 50–600 Hz range, using an 1 Hz increment. The fundamental frequency estimate is the candidate F_0 that minimizes the weighted error sum $E(F_0)$ defined as[7]

$$E(F_0) = \frac{1}{F_0} \sum_{n=1}^{N} \alpha(\nu_n) \mid S_\phi(\nu_n) - \lfloor \frac{S_\phi(\nu_n)}{F_0} + 0.5 \rfloor \, F_0 \mid \tag{5.4}$$

where ν_n is the center frequency of the nth Gabor filter in the filterbank, N is the total number of filters in the filterbank and $S_\phi(\nu_n)$ is the phase slope for the band centered at frequency ν_n. The weighting factors $\alpha(\nu_n)$ measure the relative prominence of the estimated harmonic $S_\phi(\nu_n)$. They are defined as the root mean square amplitude envelope

$$\alpha(\nu) = \left(\frac{1}{T} \int_{t_0}^{t_0+T} a^2(t,\nu)\, dt \right)^{\frac{1}{2}} \tag{5.5}$$

where t_0 and T are the start and duration of the analysis window respectively. In the error sum of Eq. (5.4), deviations of the phase slope estimate from the nearest multiple of the fundamental frequency candidate are penalized. The estimated fundamental frequency F_0 provides the best match between the short–time harmonic estimates $\{n : S_\phi(\nu_n)\}$ and the fundamental frequency multiples $\{k : kF_0\}$.

[7] $\lfloor . \rfloor$ denotes truncation of the decimal part and $\lfloor . + 0.5 \rfloor$ is the rounding operator.

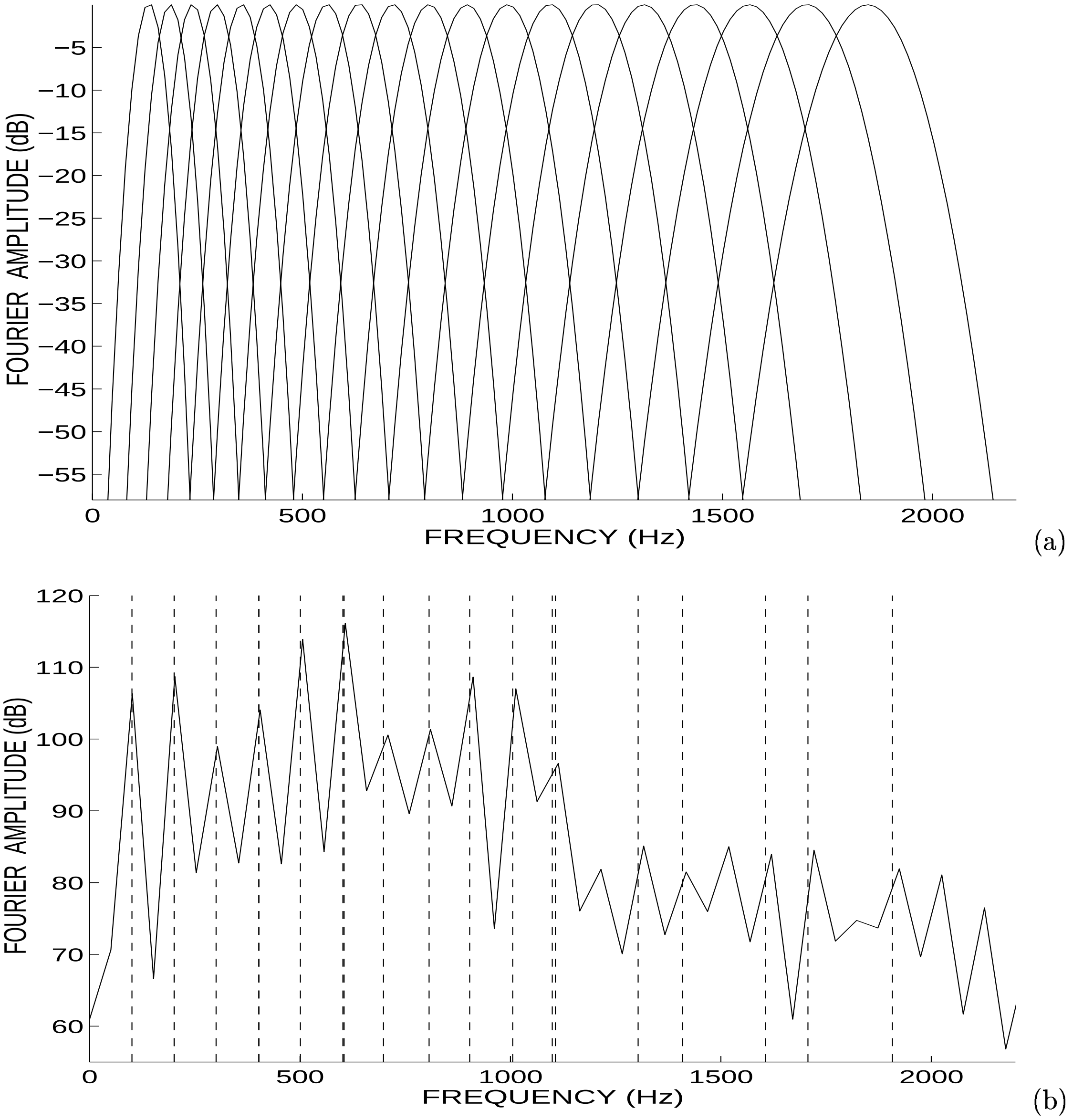


Figure 5.2: (a) The Mel–spaced "dense" Gabor filterbank used for MDA and (b) the phase slope $S_\phi(\nu)$ estimates for each frequency band ν shown superimposed on the Fourier spectrum for a 20 msecs speech frame (/aa/ from "dog").

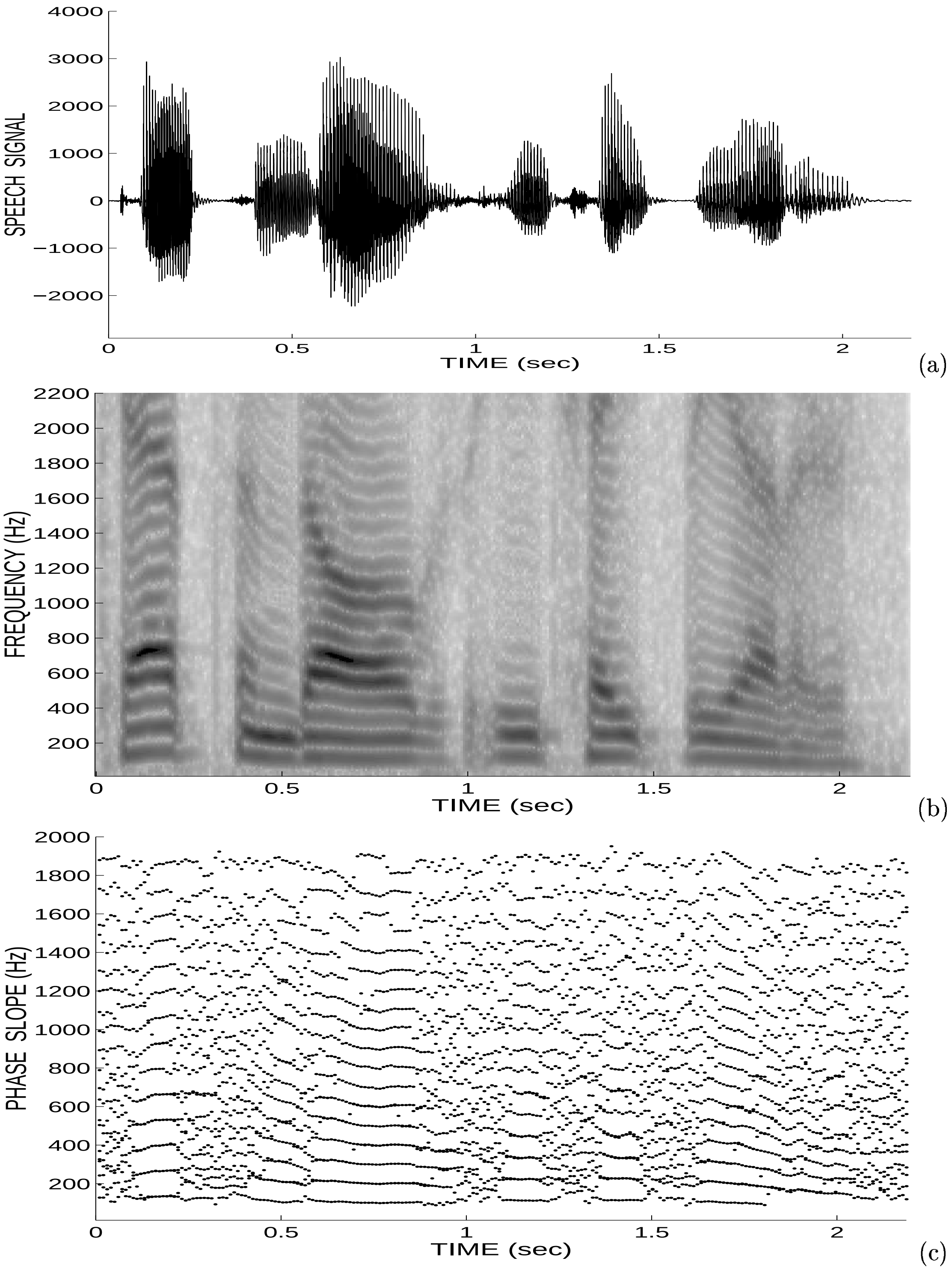


Figure 5.3: (a) Speech signal: "Cats and dogs each hate the other", (b) narrowband speech spectrogram, (c) short–time phase slope estimates for each of the 20 frequency bands vs. time (20 msecs window, 10 msecs update).

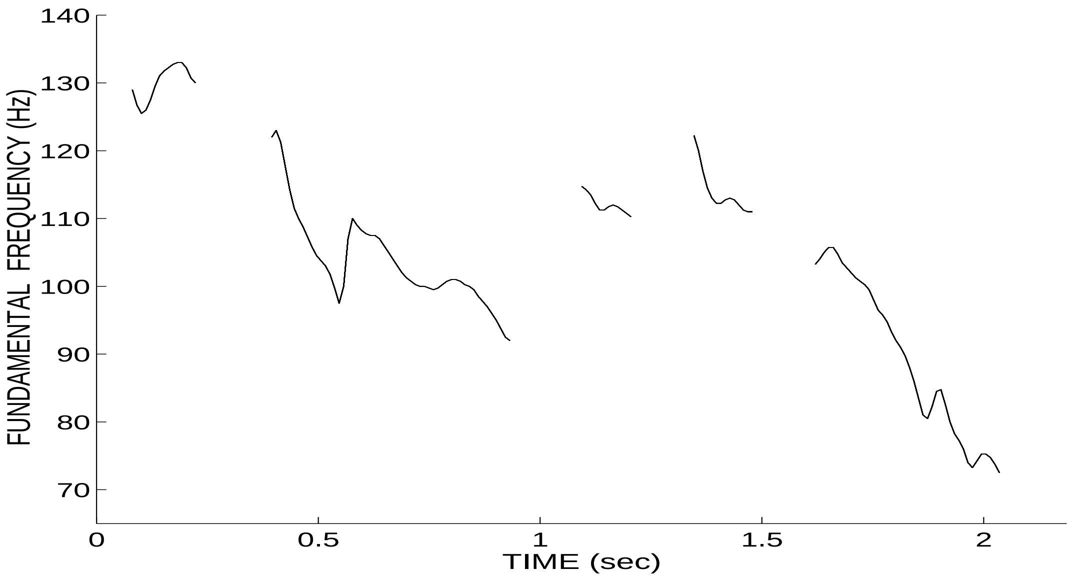


Figure 5.4: Fundamental frequency contour computed from the multiband demodulation pitch tracker for the TIMIT sentence: "Cats and dogs each hate the other" (20 msecs window, 10 msecs update).

The multiband demodulation pitch estimation algorithm produces accurate and very smooth pitch contours. The error criterion proposed in Eq. (5.4), though, sometimes suffers from "pitch halving", i.e., estimating half of the actual pitch. To correct this problem a long–time average pitch estimate is computed from 5–6 pitch periods of steady–state voicing. A Gaussian window centered on this "global" fundamental frequency estimate (with a 200 Hz bandwidth) is used to weight the error function $E(F_0)$. The Gaussian weighting improves significantly the robustness of the pitch tracker.

Due to the formulation of the pitch estimation algorithm, an explicit voiced/unvoiced decision must be made, because the algorithm assigns a pitch estimate even for unvoiced regions. Currently, a simple voicing decision is used based on the short–time "classic energy" and the first spectral moment.

For unstable voiced regions a nonlinear post–filter is used to enhance the smoothness of the fundamental frequency estimates. Regions of unstable voicing are defined as regions where the fundamental frequency estimate varies more than 5% among three or more neighboring frames. Unstable regions are smoothed by applying a 3–point binomial filter 3 times. Note that after each pass the neighboring stable regions are restored to their original values. Finally, a 3–point binomial filter can be applied on the pitch contour for increased smoothness.

In Fig. 5.4, we show an example of fundamental frequency estimation using multiband demodulation for the sentence shown in Fig. 5.3 (a). Note the smoothness and detail of the fundamental frequency contour.

Overall, the performance of the multiband demodulation pitch tracker is very good: it produces smooth estimates and does not suffer from "pitch doubling". The algorithm is conceptually simple and can be implemented in parallel for increased speed. Finally, one can refine the estimates in an iterative procedure by gradually reducing the analysis window duration, thus enhancing the pitch contour detail while preserving its smoothness (see Section 5.5).

5.4 Relation to other Pitch Estimation Methods

The sinusoidal model [62] represents the speech waveform as a superposition of sinusoids. For voiced speech the frequencies of the sinusoids are harmonically related. This relationship can be exploited to determine the fundamental frequency and the probability of voicing. In [63, 23], sinusoidal–based pitch estimation algorithms are proposed that use an error functional minimization procedure similar to Eq. (5.4) to determine the most probable fundamental frequency candidate. Clearly the sinusoidal model and the multiband demodulation pitch estimation algorithms are closely related, because in both algorithms the pitch is computed by estimating the harmonic frequencies. Yet, the MDA algorithm is different in the following important ways: (i) the harmonic estimate is computed from the slope of the phase signal in the time domain, (ii) a fixed filterbank is used for the analysis and a harmonic estimate is obtained from each frequency band (as a result some harmonics may not have a corresponding phase slope estimate, while more than one estimates may correspond to other harmonics, e.g., see Fig. 5.2 (b)), (iii) the error functionals of the MDA- and sinusoidal–based pitch trackers are different. From preliminary comparisons between the sinusoidal–based pitch estimation algorithm proposed in [23] and the MDA we have observed that the MDA produced smoother estimates and was less susceptible to "pitch doubling". Formal comparisons are currently under way. Overall, the MDA pitch estimation algorithm is attractive due to its conceptual simplicity, good performance and parallel time–domain implementation.

The human ear has often been modeled by a filterbank [40]. In [67, 68], McEachern asserts that the human ear demodulates the output of each filter and computes the time average of the frequency modulating signal (instantaneous frequency). Typically, a single prominent harmonic will fall in each frequency band and the value of this harmonic is tracked. Finally, McEachern speculates that the auditory system perceives the fundamental frequency as a weighted sum of the estimated harmonics. The MDA pitch tracker uses a multiband demodulation analysis scheme similar to the one suggested above. Clearly, the Gabor filterbank structure is different from the typical auditory filterbank and the energy separation algorithm may be very different from the demodulation approach that the human ear possibly uses. It is probable, though, that the MDA has similarities with the the way the human ear detects the fundamental frequency.

5.5 Discussion

The design of the Gabor filterbank is an important implementation issue for the MDA pitch estimation algorithm. We have observed that an adequate number of filters for robust fundamental frequency estimation is approximately 10 filters in the 0–2 kHz range. The performance of the algorithm improves as the number of filters increases, up to approximately 20 filters. Non–uniform frequency spacing of the filters is essential for robust pitch detection. As seen in Fig. 5.3 (c) the phase slope contour $S_\phi(t, \nu)$ for the band centered at frequency ν shifts between neighboring harmonics as the pitch contour evolves with time, e.g., for the fourth band S_ϕ shifts from the fourth to the third harmonic as the pitch decreases. Non–uniform spacing of the Gabor filters guarantees that such shifts will not happen simultaneously in many filters and confuse the pitch estimation procedure.

Another interesting issue is the choice of the pitch decision algorithm. We are currently using the error criterion of Eq. (5.4) weighted by a Gaussian window centered at a "global" fundamental frequency estimate. The Gaussian weighting does not allow large deviations from the average fundamental frequency and is equivalent to a continuity constraint imposed on the pitch contour. Alternatively, a global error functional can be defined for each voiced region that explicitly penalizes pitch discontinuities. Global error minimization is, in general, computationally expensive but it produces robust pitch estimates. The global error E_G to be minimized over all possible pitch paths $F_0(t)$ is defined as

$$E_G = \int_{t_1}^{t_2} E[F_0(t)]\, dt + \lambda \int_{t_1}^{t_2} \frac{d^2 F_0(t)}{dt^2}\, dt \tag{5.6}$$

for each voiced region $[t_1, t_2]$. E is the error criterion of Eq. (5.4) and λ is a scalar that weights the relative importance of the error terms. Smoother pitch contours are obtained for large values of λ. The computational efficiency of the global error minimization procedure can be increased by pruning of the less probable paths.

Finally, a multiscale approach can be used to refine the MDA pitch contour. After the pitch contour is computed at a coarse scale, the estimates are enhanced by gradually decreasing the duration of the short–time averaging window, and recomputing the phase slope S_ϕ. This can be done at a small additional computational expense since the phase signals $\phi(t)$ are already computed.[8] The window duration can be adaptively reduced to approximately 1.5 times the pitch period to produce a detailed and smooth pitch contour. The multiscale MDA approach could also improve the performance of the pitch estimation algorithm for noisy speech.

[8] This is a distinct advantage of the time–domain implementation.

Chapter 6

The AM–FM Modulation Vocoder

6.1 Introduction

Voice coding (vocoding) is one of the most important applications of speech processing. Coders can be used for both off–line storage of speech signals and for real–time communication applications, e.g., telephony and cellular telephony. The first efforts in speech coding started almost five decades ago, yet, the area is still very active today. Typically, speech coders are classified according to their information bandwidth, i.e., the number of bits used to represent a second of the speech signal. In this chapter, we will concentrate in low and mid bit rate coders from 2.4 to 9.6 kbits/sec. In order to achieve good reconstructed speech quality at such low bit rates, the structure of the speech signals must be fully exploited. The pitch periodicity and the formant structure account for most of the structure in speech. This a priori knowledge is used by most speech coding schemes that are being reviewed in the next sections.

The first step in designing a vocoder or a speech analysis–synthesis system is the choice of a speech production model.[1] In the next subsections, we review some of the most successful speech coding models, including linear prediction (see also Section 1.2) and the sinusoidal model. We conclude the section with an overview of the formant and the phase vocoders. In Section 6.2, we introduce the *AM–FM modulation analysis–synthesis system*

[1] The fact that a clear choice of a speech model does not exist today is an indication of our limited understanding of the dynamics of speech production.

that is based on the AM–FM speech model. The information signals of the vocoder are the amplitude envelope and instantaneous frequency signals of the four most prominent formant bands. The information signals are obtained from time–varying bandpass filtering followed by demodulation. In Section 6.3, we investigate efficient modeling schemes for the information signals. The perceptual importance of the amplitude envelope and the instantaneous frequency is determined qualitatively from listening tests, and serves as a guide in this modeling effort. Finally, in Section 6.4 efficient coding and quantization schemes are proposed for the *AM–FM modulation vocoder* operating in the 4.8–9.6 kbits/sec range. We conclude this chapter with informal listening tests to determine the synthetic speech quality. Overall, the AM–FM vocoder produces very natural sounding speech and can be efficiently used for various speech processing applications including vocoding and text–to–speech synthesis.

6.1.1 Vocoders Based on the Source–Linear-Filter Model

A popular speech model is the *source–linear-filter* (SLF) model, where speech is modeled as the convolution of a general excitation signal and an all–pole linear filter that represents the vocal tract (see Section 1.2). The coefficients of the linear filter are estimated through a mean square error minimization procedure, i.e., linear prediction (LP) (Eq. (1.10)). The linear prediction vocoder (LPC–10) models the excitation signal either with a pulse train for voiced sounds or with white noise for unvoiced sounds [79]. Such modeling produces synthetic speech of unnatural quality. Efficient characterization of the excitation signal is a non trivial task since one has to account not only for source variabilities (e.g., mixed voiced–unvoiced excitation, unstable voicing, double pitch), but also for the distortion introduced to the excitation signal by the deficient linear time–invariant all–pole vocal tract model.

A mathematically tractable procedure for modeling the excitation signal is the analysis–by–synthesis method, where an exhaustive ad–hoc search takes place over a set of parameters of the excitation model. The excitation is modeled as a series of pulses in *multipulse LPC vocoders* [4, 99] and/or selected from a codebook of random signals in *code–excited LPC vocoders* [96, 66]. These variants of the source–LP-filter modeling produce good speech quality at 4.8–9.6 kbits/sec [10].

The *multiband excitation vocoder* is an extension of the source all–pole filter vocoder.

The frequency axis is divided into fixed frequency bands and each band is allowed a separate voiced/unvoiced decision. This results in a different excitation signal for each band, the choice being between a train of pulses or white noise. The mixed excitation scheme improves synthetic speech quality [25].

6.1.2 Vocoders Based on the Sinusoidal Model

Alternatively, one can model speech as a sum of sinusoids, one for each speech harmonic [62]. To estimate the parameters of the *sinusoidal model*, the amplitudes and the frequencies of the sinusoids are assumed constant over a short speech segment. The frequencies of the sinusoids can be estimated either from peak picking of the short–time Fourier transform [62], or from analysis–by–synthesis [23]. For efficient coding, the sinusoids are assumed to be harmonically related for voiced speech. To resynthesize the speech signal the frequencies of the sinusoids between neighboring speech frames are carefully matched to create smooth frequency contours. Finally, the short–time amplitudes and frequencies are interpolated and used to reconstruct the sinusoids [64].

The success of the sinusoidal vocoder is mainly due to the efficient modeling of the quasi–periodicity in the speech signal and to the "sinusoidal interpolation" synthesis scheme. Overall, the sinusoidal model is very popular, and has been successfully applied to speech coding in the 2.4–4.8 kbits/sec range [64], time–scale and pitch modification [90], and pitch tracking [63].

6.1.3 Formant Vocoders

In *cascade formant vocoders*, the excitation signal is convolved in series with the impulse response of three to five resonators, one for each speech formant. Each resonator is typically modeled as a second–order linear system with stepwise constant parameters. *Parallel formant vocoders* model the speech signal as the superposition of the responses of the resonators to the excitation signal [92].

The series implementation does not allow independent control of the amplitude of each formant, and is equivalent to a source–all-pole filter model. In general, cascade formant vocoders produce more natural sounding speech than LP vocoders.[2] The parallel implementation allows distinct excitation inputs and, as a result, independent amplitude control for each formant. This is a more general model that can accurately represent complex speech spectra, e.g., fricatives [31, 41].

Formant vocoders produce good speech quality at very low bit rates (1.0–2.0 kbits/sec), yet, they are not very robust, since estimating the parameters of the coder is an open–loop procedure. As a result they are mainly used today as speech synthesizers in text–to–speech systems where the formant parameters are produced by synthesis rules [41, 43, 44]. By carefully and painstakingly adjusting the parameters of the formant synthesizer speech of very good quality can be produced.[3] Recently, preliminary efforts have been made to incorporate some of the effects of nonlinearities during speech production in formant synthesizers. This is achieved by either modifying the excitation characteristics (glottal pulse shape) [11, 2] or by adding simple frequency modulation patterns to the formant frequencies to improve speech naturalness [44, 2].

6.1.4 Phase Vocoder

Fourier phase information in a speech signal is much less important than amplitude information, especially for frequencies over 1 kHz. The *phase vocoder* exploits this fact efficiently. The speech signal is filtered through a bank of fixed (in frequency) bandpass filters and, then, each band is demodulated to the amplitude envelope and phase components. The amplitude envelope and instantaneous frequency signals are then decimated and coded [18, 92]. Note that since the phase information is less important perceptually than the amplitude envelope information, the instantaneous frequency signal is sampled very coarsely. If the filterbank is not designed carefully, though, decimation of the instantaneous frequency can introduce very noticeable distortion in the synthetic speech signal [19, 20].

Alternatively, one may use a time–varying filterbank so that each band is centered at the pitch harmonics for voiced speech. The "pitch–tracking" phase vocoder was proposed by Malah in [51].

[2] The continuity in the formant tracks is in accordance with the "smooth" movements of the articulators.

[3] Unfortunately, due to our limited understanding of the dynamics of speech production, the fine tuning of a formant synthesizer is still today more of an art rather than a science.

6.2 The AM–FM Modulation Analysis–Synthesis System

The AM–FM modulation model was introduced in Section 1.4 and represents a speech resonance as a signal with a combined amplitude and frequency modulation. Speech is modeled as the sum of three to four resonance signals. We introduce next an application of this model to speech coding.

The AM–FM modulation analysis–synthesis system extracts three or four time–varying *formant bands* $r_k(t)$ from the spectrum by filtering the speech signal $s(n)$ along the formant tracks. The formant tracks are obtained from the multiband demodulation (MDA) formant tracking algorithm introduced in Section 4.4. The filtering is performed by a bank of Gabor filters with time–varying center frequencies that follow the formant tracks. The MDA formant tracker bandwidth estimates determine the bandwidth of the Gabor filters.[4] Next, the formant bands are demodulated to the amplitude envelope $a_k(t)$ and the instantaneous frequency $f_k(t)$ signals using the energy separation algorithm (ESA). The amplitude envelope and instantaneous frequency are the information signals of the AM–FM modulation vocoder to be modeled and coded in the following sections. The information signals are decimated by a factor of 20:1 (original sampling frequency at 16 kHz) before being modeled or coded. Typical bandwidths for the information signals are in the 600–1000 Hz range.

In Fig. 6.4 (b) and (c), we display the amplitude envelope (for F1, F2) for an 80 msecs segment of the phoneme /ow/ from "zero". The corresponding instantaneous frequency signals for F1, F2 are shown in Fig. 6.5. The information signals are decimated to a 600 Hz bandwidth. Note that apart from the evident quasi–periodic structure, the information signals contain a considerable amount of amplitude and frequency modulations. We have observed that when the modulations are removed through excessive decimation, the naturalness of the synthesized speech deteriorates considerably.

To synthesize the speech signal, the phase is obtained as the running integral of the instantaneous frequency, and the formant bands $\hat{r}_k(t)$ are reconstructed from the amplitude and phase signals. When the amplitude envelope and instantaneous frequency are not coded, there is no band reconstruction error, i.e., $\hat{r}_k(t) = r_k(t)$. The synthetic speech signal

[4] In the AM–FM modulation vocoder, the bandwidths of the Gabor filter are set to 300 Hz for the first and to 400 Hz for higher formants to simplify the amplitude coding scheme.

$\hat{s}(t)$ is obtained as the sum of the reconstructed formant bands. The block diagram of the AM–FM modulation analysis–synthesis system is shown in Fig. 6.1.

The AM–FM modulation vocoder models individually each formant band and is, in that sense, similar to the parallel formant vocoder. Yet, instead of modeling the signal as the output of a source–filter system as most vocoders do, each formant band is demodulated into its amplitude and frequency components. By retaining the excitation–vocal tract coupling the hard decomposition problem is avoided temporarily; this offers more freedom to explore nonlinear speech production phenomena not modeled by the source–linear-filter model. Further, the modulation structure has the physical interpretation of being the amplitude and frequency modulation existent in the speech resonances. Finally, the perceptual importance of amplitude and frequency modulations in speech formant bands can be quantified once the information signals are modeled (see Section 6.3).

The AM–FM vocoder is related to the phase and the parallel formant vocoders. Both the AM–FM and the formant vocoders model the speech signal as a superposition of formant signals. The important difference is that instead of making the quasi–stationarity assumption, the AM–FM vocoder describes each formant resonance by two signals (amplitude and frequency) that are allowed to vary instantaneously with time. As a result, the AM–FM vocoder breaks free of the source–linear-filter assumption and can efficiently represent and model any general speech resonance signal. The representation of a speech band by the amplitude envelope and instantaneous frequency signals is common ground between the phase and AM–FM vocoder. Yet, the AM–FM vocoder uses a time–varying filterbank following the formant tracks to extract the formant bands (instead of a fixed one). Furthermore, as we shall see, the modeling and coding of the information signals is performed in a novel way. The similarity between formants and the quasi–periodicity of the information signals is exploited in a multipulse coding scheme that is proposed for the amplitude envelope signals in Section 6.3. This leads to significant bandwidth savings.

The amplitude envelope $a(t)$ and instantaneous frequency $f(t)$ signals of a speech resonance have a specific structure. Next, we elaborate on the structure and the perceptual importance of the formant amplitude and frequency signals. Our end goal is to obtain a detailed and concise representation of the information signals.

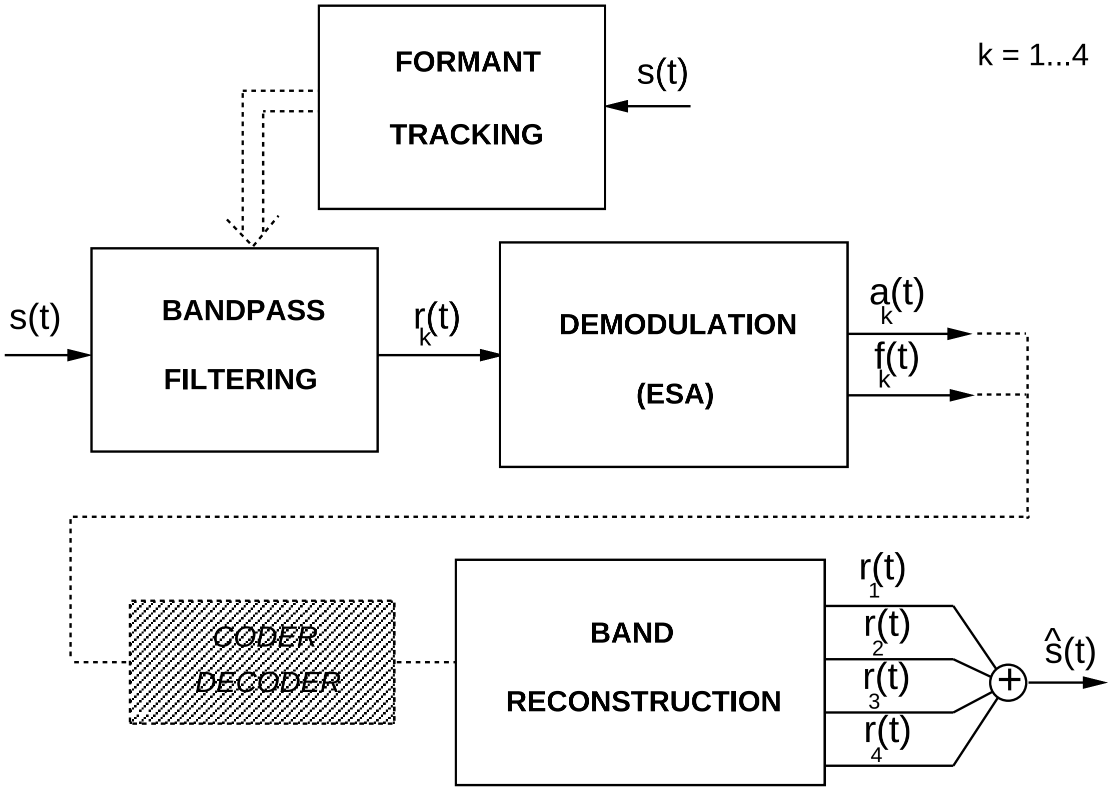


Figure 6.1: The block diagram of the AM–FM modulation analysis–synthesis system. Four formant bands $r_k(t)$ are extracted from the speech signal $s(t)$ through bandpass filtering. Each formant band $r_k(t)$ is demodulated to amplitude envelope $a_k(t)$ and instantaneous frequency $f_k(t)$ signals, $k = 1, 2, 3, 4$. Speech is synthesized by adding together the reconstructed formant bands $\hat{r}_k(t) = r_k(t)$.

6.2.1 The Amplitude Envelope Signal

LPC and formant vocoders typically model the amplitude envelope of a speech resonance as an exponentially decaying signal for a pitch period of voiced speech. In reality, as shown in Fig. 6.4 (b), the amplitude envelope signal often presents secondary "energy" pulses and/or changing rate of energy dissipation inside a pitch period, i.e., bandwidth modulation. The modulation patterns are mainly due to secondary excitations after glottal closure [30], flow instabilities in the vocal tract [103], and, in general, nonlinear interaction between the vocal tract and the glottis.

For voiced speech, the amplitude envelope mainly conveys information about the source of voicing. The excitation timing information is perceptually crucial for voiced speech because it determines the possible periodic structure of the speech resonance signal. In addition, amplitude modulation in a single pitch period has been found to be perceptually important. We have observed that when the amplitude modulations are removed from the speech resonances the naturalness of the synthetic speech degrades significantly. Thus, the fine structure of the amplitude envelope has to be modeled accurately for voiced speech.

For unvoiced speech the detailed shape of the amplitude envelope is of little perceptual importance. More important is the overall shape of the amplitude envelope that determines the amplitude of the formant peak.

The effects of the Gabor bandpass filter (used during the analysis stage of the vocoder) on the amplitude envelope signal were discussed in Section 3.4.2. The estimated amplitude envelope signal $\tilde{a}(t)$ is approximately a lowpass filtered version of the actual one $a(t)$, as seen from Eq. (3.28). Specifically, for speech resonances: $\tilde{a}(t) \approx a(t) * h_\ell(t)$, where $h_\ell(t)$ is the Gaussian (lowpass Gabor) filter. The approximation is valid provided that the deviation of the instantaneous frequency from the formant frequency is small. We have observed that for speech resonances the above stated approximation is a reasonable one.

6.2.2 The Instantaneous Frequency Signal

In source–linear-filter based vocoders, the formant frequency is assumed to be constant during a short speech segment. As a result, the corresponding instantaneous frequency of the synthetic speech resonance is constant everywhere and equal to the formant frequency.

Around excitation instants, though, there is a spike with value equal to the phase difference between the decaying resonance signal of the previous and the current pitch period. The instantaneous frequency spikes and the phase discontinuities around excitation instants can be seen in Fig. 5.1 (c) and (d) for a synthetic speech resonance synthesized from a linear second–order system. The area under the instantaneous frequency spikes is important because it affects the phase relation between consecutive pitch periods as we discuss next.

The instantaneous frequency signal contains information about the average formant frequency value. Clearly, the formant frequencies and, especially, the time variation of the formants is perceptually very important information [42, 45]. Further, to achieve good quality for voiced speech the correct phase relation between consecutive pitch periods is also essential. When the phase modeling is wrong, the synthetic speech resonance waveform becomes aperiodic and speech sounds dissonant. This is especially true for formants below 1 kHz (i.e., mainly F1).[5] To guarantee the naturalness of voiced speech the phase discontinuities (or equivalently the instantaneous frequency spikes) must be modeled accurately. For unvoiced speech the phase information is perceptually less important than for voiced speech. Formant bandwidth information is conveyed through the value of phase discontinuities.

Apart from the slow–varying average formant value and the phase at excitation instants the instantaneous frequency contains additional modulation information. Specifically, the instantaneous frequency signal for the open and the closed phase often takes different values as shown in Fig. 6.5. We have observed that such modulations are perceptually important only for low formants (mainly F1).[6] The instantaneous frequency modulations for F1 are often important (frequency deviations up to 150 Hz has been recorded) and have to be modeled.

Finally, the main effect of the Gabor bandpass filter on the instantaneous frequency signal is to lowpass filter the spikes occurring around excitation instants, e.g., compare

[5] We have observed that when the ESA is used for demodulation the size of the instantaneous frequency spikes is often estimated incorrectly, especially, for formants in the 0–500 Hz frequency range. This can be simply corrected through the model proposed for the instantaneous frequency in Section 6.3.2. Alternatively, the Hilbert transform demodulation can be used for F1.

[6] This is partly due to the reduced frequency discrimination of the human ear for higher frequencies [18].

Figs. 3.4 and 3.6. For a constant amplitude envelope (no AM) it follows from Eq. (3.29) that the instantaneous frequency spikes become Gaussians, i.e., they are filtered by a lowpass Gabor filter. For speech, the filtered spikes are still bell–shaped but they are no longer symmetric because the formant amplitude envelope is time–varying.

6.3 Amplitude Envelope and Inst. Frequency Modeling

Next, we attempt to model the amplitude envelope $a(t)$ and instantaneous frequency $f(t)$ signals so that they can be efficiently coded and quantized in Section 6.4. As discussed above, the formant resonance information signals $a(t)$ and $f(t)$ have a specific structure that can be efficiently exploited to reduce the information bandwidth of the vocoder; e.g., for voiced speech the amplitude envelope for all formants looks similar. The (relative and absolute) perceptual importance of $|a(t)|$ and $f(t)$ serves as a guide in this modeling effort.

6.3.1 Modeling the Amplitude Envelope

Adaptive–Differential Pulse Code Modulation (ADPCM) [35] has been used successfully in the past to code the amplitude envelope signals in a phase vocoder application [20]. Yet, as we have discussed, for the AM–FM modulation vocoder the amplitude envelope signals of different formants are highly correlated for voiced speech and have a specific structure. In order to exploit these features, we use a multipulse source–filter model for the amplitude envelope. We expect the excitation signals for the amplitude envelopes of different formant bands to be coupled for voiced speech and loosely coupled for unvoiced speech.

The linear prediction–multipulse coder

In linear prediction (LP) multipulse speech coders, the excitation signal $u(n)$ consists of K pulses (per speech analysis frame), fully defined by their amplitudes b_k and positions n_k [4, 99]

$$u(n) = \sum_{k=1}^{K} b_k \, \delta(n - n_k) \tag{6.1}$$

where $\delta(n)$ is the Kronecker delta. The excitation signal $u(n)$ is convolved with the impulse response $w(n)$ of an all–pole filter (that models the vocal tract) to produce the synthetic

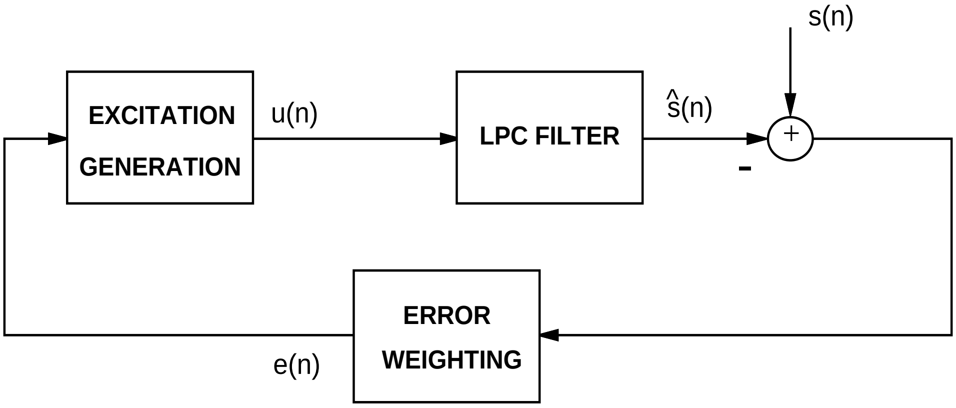


Figure 6.2: Analysis–by–synthesis multipulse loop for speech signals $s(n)$.

signal $\hat{s}(n)$, i.e.,[7]

$$\hat{s}(n) = u(n) * w(n) \tag{6.2}$$

First, the parameters of the filter $w(n)$ are estimated by linear prediction. Then the excitation pulse positions n_k and amplitudes b_k are obtained by minimizing the mean square modeling error E

$$E = \sum_{n=1}^{N} e(n)^2 = \sum_{n=1}^{N} [s(n) - \hat{s}(n)]^2 \tag{6.3}$$

where N is the size of the speech analysis frame. Note that often the modeling error $s(n) - \hat{s}(n)$ is convolved with a perceptual filter that weights more the spectral valleys contribution to the error (see Fig. 6.2) [4]. The positions n_k of the excitation pulses are determined sequentially (one at a time), through an exhaustive analysis–by–synthesis search, since no closed form solution exists for the minimization of E over n_k. The amplitudes b_k are computed from the system of linear equations obtained from minimizing E. The multipulse analysis–by–synthesis loop is outlined in Fig. 6.2.

Modifications to the multipulse coder

As discussed in Section 6.2, the formant amplitude envelope signals are highly correlated for voiced speech; one would expect this similarity to be mirrored in the excitation signals. When applying the multipulse model of Fig. 6.2 directly on the amplitude envelope of the

[7] For simplicity, the contribution from the previous analysis frame is ignored here.

formant bands the correlation between $u(n)$ of different formants is small and, as a result, excessive bandwidth is needed to code the excitation signals. This is due to: (a) the inability of the linear predictor to extract the important information from the amplitude envelope signals and (b) the distortion introduced to the amplitude envelope signal from the Gabor bandpass filtering procedure. Next, we attempt to remedy these two shortcomings of the multipulse coding approach.

Often the amplitude envelope signal $a(n)$ is corrupted either by noise or by the neighboring formants (see Section 3.4.3). In both cases, the linear predictor performs poorly and the resulting excitation signal is noisy. To overcome this problem, we model the amplitude envelope as the superposition of shifted versions of the impulse response of a critically damped baseband second–order system. This mono-parametric family of waveforms was found to be a reasonable first–order approximation to the amplitude envelope of real speech resonances for both the attack and the (exponential) decay portions of the signal. Further, the modeling of the amplitude envelope is greatly simplified because only a single parameter has to be estimated. Finally, by incorporating this prior knowledge about the typical shape of the amplitude envelope pulses, the multipulse excitation signal $u(n)$ becomes more robust to adverse degradations of the amplitude envelope. The frequency response of this simple second–order single–parameter linear predictor is

$$G(z) = \frac{1 + c_1 + c_2}{1 + c_1 z^{-1} + c_2 z^{-2}}, \quad c_1 = -2\exp(-\pi B/F_s), \quad c_2 = \exp(-2\pi B/F_s) \qquad (6.4)$$

where B determines the rate of decay of the amplitude envelope signal (B is equivalent to the formant bandwidth), and F_s is the sampling frequency. Note that a single parameter B has to be estimated for each speech analysis frame to fully determine the predictor.[8] The rate of decay B (or formant bandwidth) parameter is computed from analysis–by–synthesis. We have found this method to be more robust that an open loop estimation, while the increase in computational complexity is modest.[9]

As discussed in Section 6.2.1 the amplitude envelope estimate is (approximately) a

[8] Alternatively, a first–order linear predictor can be used to model the exponentially decaying amplitude envelope. We have observed, though, that the impulse response of the second–order system resembles more closely the actual amplitude envelope of speech resonances, especially, for the "attack" region.

[9] The bandwidth parameter B is assumed to be slow–varying and for each frame 3–5 candidates in the neighborhood of the previous bandwidth estimate are tested.

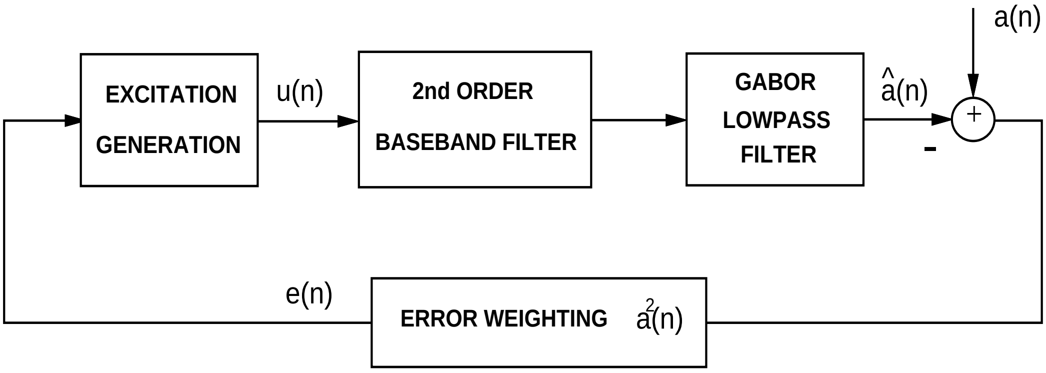


Figure 6.3: Analysis–by–synthesis multipulse loop for the amplitude envelope signal $a(n)$ using a critically damped baseband second–order filter.

lowpass filtered version of the actual envelope. To compensate for the effects of filtering a Gabor lowpass filter $h_l(t)$ (Gaussian) with bandwidth equal to that of the Gabor bandpass filter used to extract the formant band, is included in the analysis–by–synthesis scheme.

The modified multipulse model for amplitude envelope signals is shown in Fig. 6.3. The modeling error is weighted by the square of the amplitude envelope. The weighting guarantees better modeling of the envelope maxima which are perceptually very important (envelope maxima convey information about the excitation instants).

Due to these modifications to the multipulse algorithm the amplitude envelope of the formant bands can be efficiently coded, since the correlation among excitations for different formants is now evident. In Fig. 6.4, we display the amplitude envelope (decimated to a 600 Hz bandwidth) and its excitation signals for F1, F2 for the phoneme /ow/ of the word “zero”. The excitation signals are computed from the multipulse analysis–by–synthesis loop of Fig. 6.3. A maximum of six excitation pulses are assigned to a 25 msecs segment. Note that the positions of the primary pulses for F1 and F2 align. One to three pulses (typically two) per pitch period are needed to accurately model the amplitude envelope signal, as opposed to four to six pulses needed in a speech multipulse scheme [99]. The small number of pulses needed to model $a(t)$ is due to the fact that the AM–FM modulation vocoder explicitly decouples the formants bands.

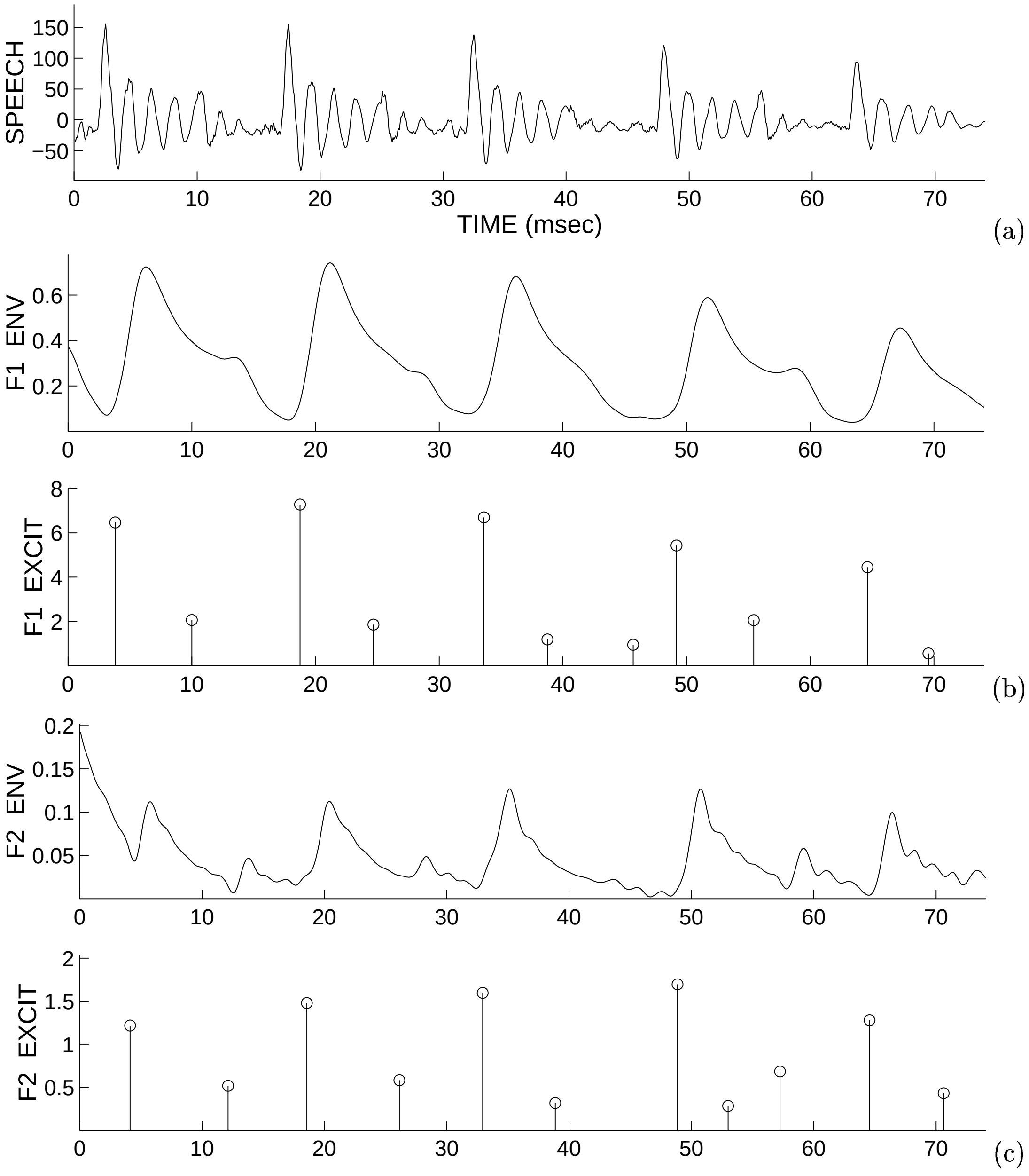


Figure 6.4: (a) Speech signal, phoneme /ow/ from “zero”. Amplitude envelope and multipulse excitation signal for the first (b) and second resonance (c).

Overall, the modified equations for the multipulse coder are

$$u(n) = \sum_{k=1}^{K} b_k \, \delta(n - n_k) \tag{6.5}$$

$$\hat{a}(n) = u(n) * w(n) = u(n) * [g(n) * h_l(n)] \tag{6.6}$$

$$E = \sum_{n=1}^{N} e(n)^2 = \sum_{n=1}^{N} [a(n) - \hat{a}(n)]^2 \, a^2(n) \tag{6.7}$$

where $g(n)$ is the impulse response of the system with frequency response $G(z)$ defined in Eq. (6.4), and $h_l(n)$ is the sampled Gaussian (Gabor lowpass filter). The excitation pulse positions p_k and the bandwidth parameter B of $G(z)$ are obtained from the analysis–by–synthesis loop. Finally, the amplitudes of the excitation pulses b_k are obtained from the following system of linear equations

$$\sum_{k=1}^{K} b_k \, [\sum_{n=1}^{N} w(n - n_k) \, w(n - n_l) \, a^2(n)] = \sum_{n=1}^{N} w(n - n_l) \, a^3(n), \quad l = 1, 2 ... K \tag{6.8}$$

where K is the number of pulses in the analysis frame of length N. Note that since the amplitude envelope $a(n)$ is always positive, pulses with negative amplitudes are not allowed in the excitation signal.

The quality of the reconstructed envelope modeled with two pulses per pitch period was tested against the original one using the AM–FM modulation analysis–synthesis system. The synthetic speech produced was almost indistinguishable from the reconstructed (sum of formant bands) signal.

An interesting enhancement to the amplitude envelope model is to allow the formant bandwidth parameter B (of the second–order predictor) to vary inside a pitch period. A simple way to incorporate bandwidth modulation in the vocoder is by using the frequency modulation model introduced in the next section. Specifically, the bandwidth difference between the open and the closed phases can be computed from the formant frequency deviation between the two phases of voicing.[10]

6.3.2 Modeling the Instantaneous Frequency

As discussed in Section 6.2.2, the instantaneous frequency signals contain information about the formant tracks, the phase around excitation instants, and the instantaneous frequency

[10] The resonant frequency of a second–order linear system is equal to $\sqrt{\omega^2 - \Gamma^2}$ where ω is the natural frequency and Γ is the damping coefficient.

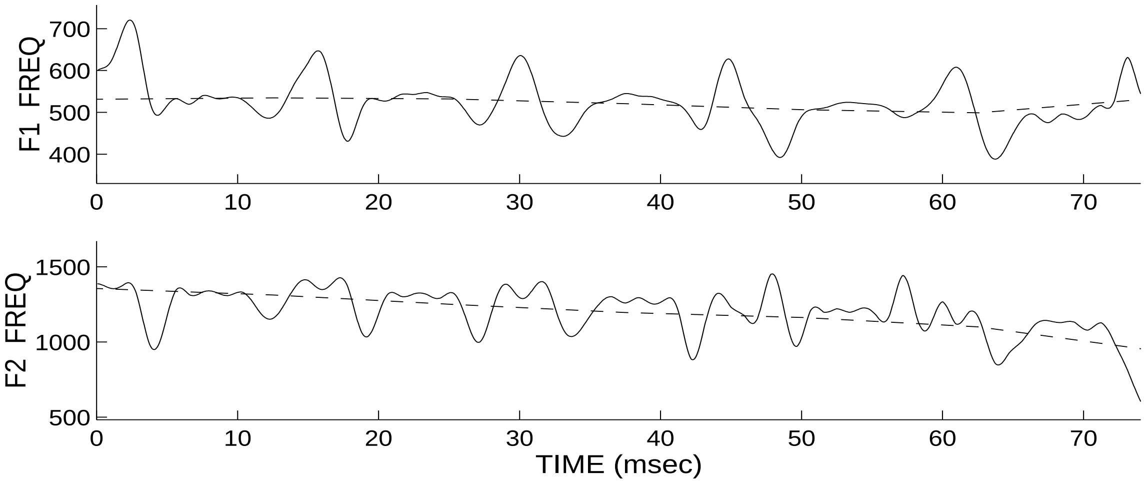


Figure 6.5: Instantaneous frequency signals and average formant tracks for F1, F2.

modulations. We attempt to efficiently model the instantaneous frequency signal without losing any of the aforementioned information.

In Fig. 6.5, the instantaneous frequency signals are shown for the first and second speech resonances (F1, F2). The average formant frequency (formant tracks) are shown as dashed lines. Note the different values that the instantaneous frequency takes for the open and the closed phases, especially for the first formant.

The formant tracks are computed from the F_w estimate defined in Eq. (4.6). Further, an additional frequency modulation component around the formant frequency is allowed. The frequency modulation component is modeled as a piece–wise linear deviation from the formant frequency. The magnitude of the frequency deviation is computed separately for the open and the closed phases. To define the open/closed phase the excitation signal obtained from the amplitude envelope multipulse model is used. Specifically, the closed phase is assumed to be the interval between the primary and the secondary excitation pulse in each pitch period.[11] A short–time estimate of the instantaneous frequency during the open and the closed phases is then computed. In Fig. 6.6 (b), we show typical short–time deviations of the instantaneous frequency from the formant track (solid line) for the open/closed phase (for the word "non"). Note that differences up to 60 Hz can be observed. In brief, the

[11] The algorithm used for grouping the excitation pulses into primary and secondary is explained in Section 6.4.1.

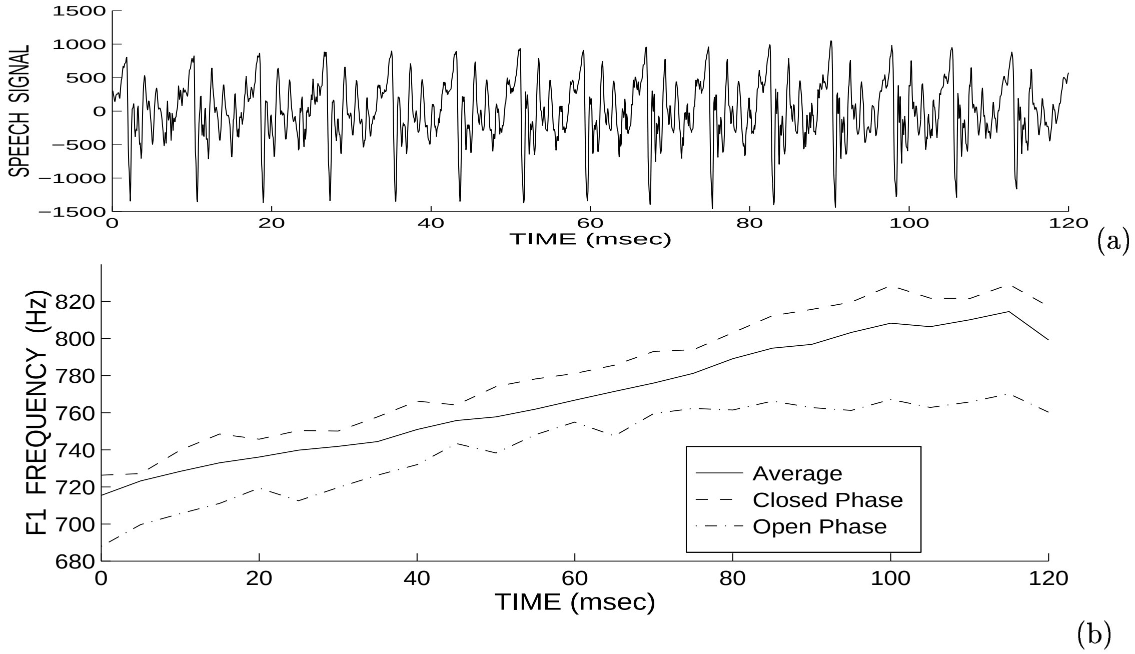


Figure 6.6: (a) Speech signal "non" and (b) F1 average formant frequency for open and closed phase (solid), for closed phase (dashed), for open phase (dashed–dotted); 15 msecs averaging window, 5 msecs update.

instantaneous frequency is modeled as a slow–varying formant frequency signal with the appropriate frequency deviation added during the open/closed phase.

Finally, to guarantee the correct phase relation between consecutive pitch periods, a smooth phase discontinuity is allowed at the excitation instants. Specifically, a scaled Gaussian is added at the excitation instants to the instantaneous frequency signal to preserve the correct phase for voiced speech.[12] The area under the Gaussian is equal to the phase discontinuity that we want to impose on the signal at that excitation instant.

6.4 Coding and Quantization

In this section, an efficient coding and quantization scheme is proposed for the multipulse modeled amplitude envelope and instantaneous frequency signals. The coding algorithm exploits the quasi–periodicity of the amplitude excitation signals. A pitch estimate is com-

[12] In reality, a spike is added to the instantaneous frequency signal and, then, the signal is filtered by a lowpass Gabor filter.

puted from the positions of the primary excitation pulses and used to efficiently code the excitation signal. Also efficient use is made of the time alignment of the excitation pulses for different formant bands to further reduce the information bandwidth. The proposed algorithms lead to an AM–FM coder operating in the 4.8 to 9.6 kbits/sec range.

6.4.1 Amplitude Envelope Coding and Quantization

In a multipulse coding scheme, the positions of the excitation pulses directly affect the (possible) quasi–periodic structure of the synthesized speech. As a result, the excitation pulse positions n_k have to be coded accurately to obtain good synthetic speech quality.[13] Next, based on the structure of the excitation signals for the amplitude envelope of different formant bands, we propose a coding scheme that results in significant bandwidth savings.

As discussed in Section 6.3, due to the decoupling of the formant bands inherent in the design of the AM–FM modulation vocoder, the amplitude envelope excitation signal has a much simpler form than the typical excitation signal of a speech multipulse scheme. It can be inferred from Fig. 6.4 that for voiced speech two to three pulses per pitch period are enough to efficiently model the amplitude envelope signal. In a pitch period, the first pulse corresponds to the primary excitation instant, while the rest model secondary excitations and nonlinear phenomena. The distance between two consecutive primary pulses is a raw estimate of the fundamental frequency. Since the fundamental frequency is a slow–varying function of time, the primary excitation pulse positions can be coded efficiently. Further, the fact that the primary pulses for different formant bands of voiced speech are aligned in time leads to additional bandwidth savings. Finally, secondary pulse positions can be coded relatively to the primary ones.

Next, we propose an algorithm for grouping the excitation pulses in the primary and secondary subgroups. The primary pulses for the first formant are determined as follows: (a) The steady state voiced regions are determined as regions where the distance between the positions of neighboring (prominent) pulses is approximately constant; an average pitch estimate is computed for each voiced region. (b) From a primary pulse located in the center

[13] Typically, a large part of the information bandwidth is assigned to n_k. The excessive bandwidth needed to code the excitation pulse positions lead to the introduction of the regular pulse excitation vocoder, where excitation pulses are uniformly spaced in time [49].

of each voiced region we search for the next primary pulse location using the average pitch estimate. A triangular window centered on the most probable primary pulse location (one pitch period from the current pulse) is used to weight the excitation. The largest weighted pulse is selected as the current primary pulse, and the pitch estimate is updated according to the distance between the current and previous pulse. The search begins from the middle of each voiced region and propagates both forward and backward in time, until the whole region of support of the signal is covered.

In Fig. 6.7 (b), we show the primary excitation pulses for the word “zero”. The distance between consecutive F1 primary pulses is shown in Fig. 6.8. Note that for voiced speech a (noisy) pitch contour is obtained, while for unvoiced regions the distance between consecutive primary pulses is random.

The locations of the primary pulses for the second and higher formant bands are modeled as perturbations around the first formant primary pulse locations. The F2 primary excitation pulses are shown in Fig. 6.7 (b). Note how the primary pulses for F1 and F2 align for voiced speech. The positions of the primary F2 pulses relative to the corresponding F1 primary pulse position are shown in Fig. 6.8.

Finally, one secondary pulse is selected in between two primary pulses. This leads to an excitation signal with two pulses per pitch period. The secondary excitation pulses for “zero” are shown in Fig. 6.7 (c).

To efficiently code and quantize the excitation positions, the contours in Fig 6.8 are modeled as the sum of a slow–varying smooth contour and a white noise component. The slow–varying amplitude of the white noise process is coded along with the slow–varying average value. This coding scheme leads to a 1 to 2 kbits/sec bandwidth for the excitation pulse position information (for corresponding vocoder information bandwidth ranges between 4.8 to 9.6 kbits/sec).

Once the excitation pulses are grouped to primary and secondary, the coding and quantization of the amplitude of the pulses is greatly simplified. In Fig. 6.4, we show that the amplitude variations between neighboring pulses are small. The smooth amplitude contours for the primary and secondary pulses are separately decimated and coded using pulse code modulation (PCM). Typically 2.5 to 4 kbits/sec are assigned to excitation pulse amplitude

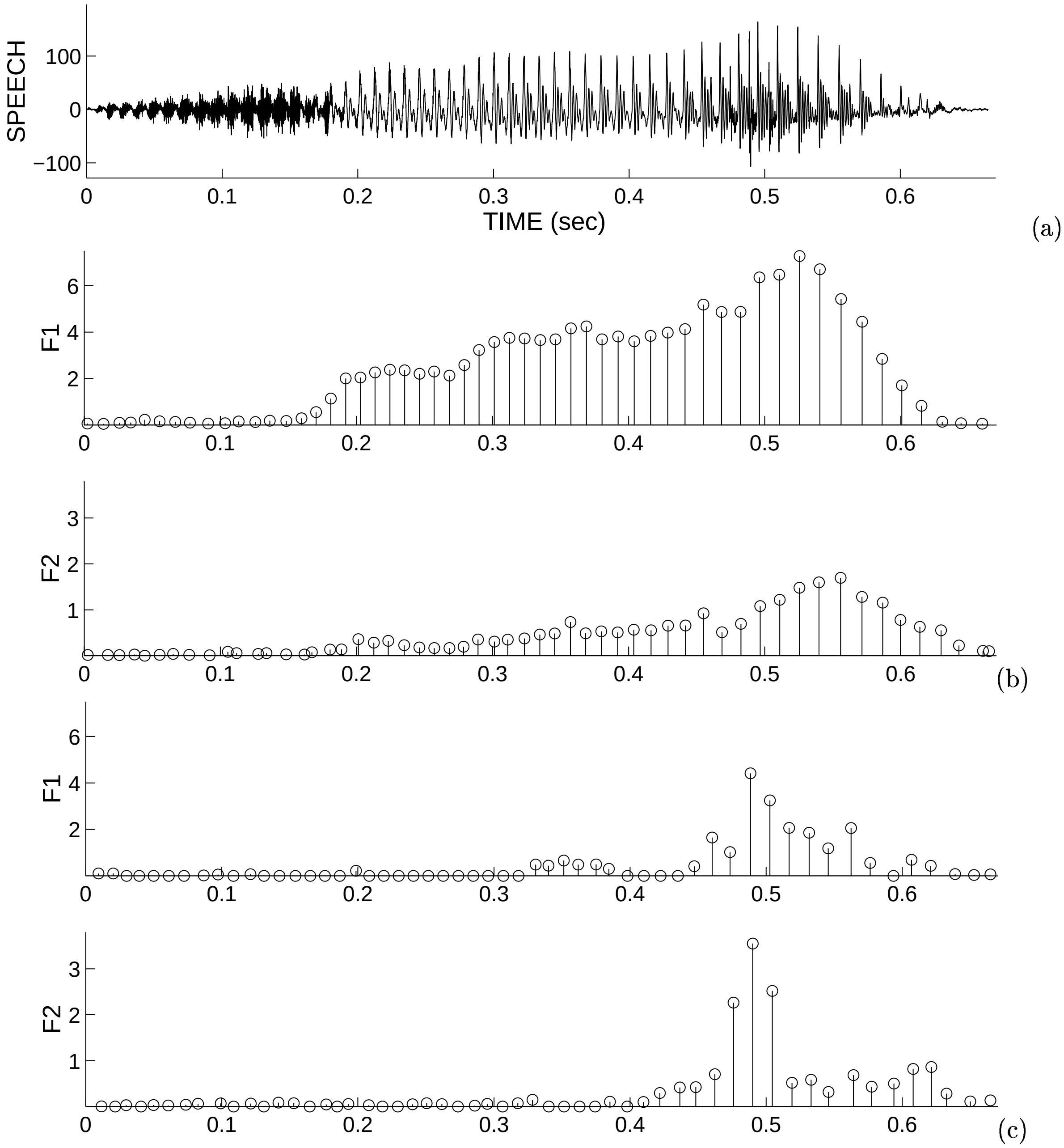


Figure 6.7: (a) Speech signal “zero”, (b) Primary excitation pulses for F1, F2, (c) Secondary excitation pulses.

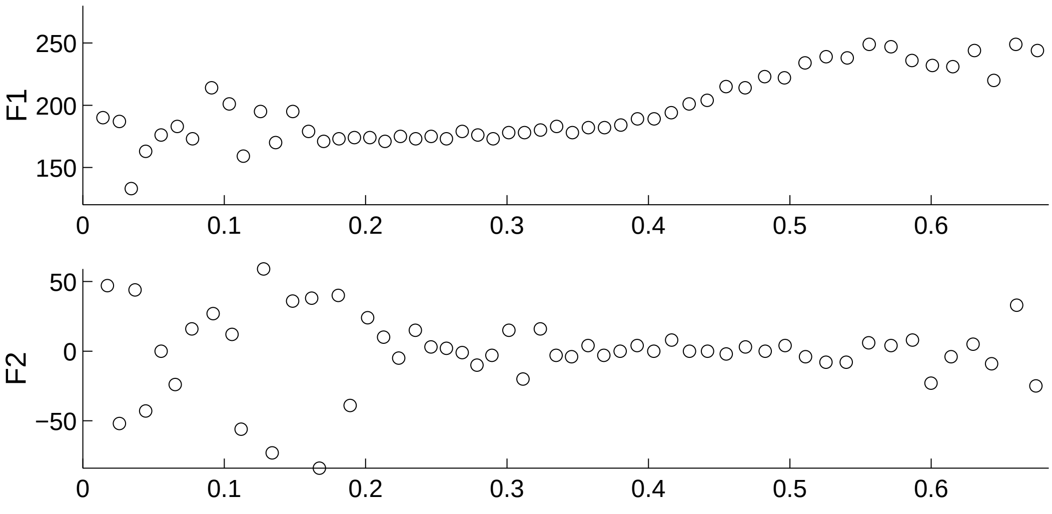


Figure 6.8: Primary pulse position coding.The distance among consecutive primary pulses of F1. The distance between primary pulses of F2 and corresponding (tied) F1 primary pulses.

coding.

6.4.2 Instantaneous Frequency Coding and Quantization

The formant contours are decimated (to 40 frames/sec) and quantized using PCM. This accounts for most of the bandwidth used to code the instantaneous frequency signals (typically 1 to 1.5 kbits/sec). Further, the value of the phase (modulo 2π) at excitation instants is coded coarsely for the first formant. Finally, the short–time amount of frequency modulation is coded for F1 only.[14] Typically, 0.5 kbits/sec are allotted to coding the modulation information for the instantaneous frequency signals.

The instantaneous frequency is reconstructed as follows: (a) the instantaneous frequency signal is set equal to the short–time formant frequency, (b) frequency modulations around the formant frequency are added, (c) bell–shaped curves are added at excitation instances so that the phase is correct (see Section 6.3.2). Note that for F2–F4 where no fine scale phase information is available, the "minimum phase" assumption is used for voiced speech and the "random phase" assumption is used for unvoiced speech [64].

[14] We have found that the detailed phase modeling makes a significant perceptual difference mainly for the first formant.

6.5 Performance

In Fig. 6.9, we display the different stages of modeling/coding of the AM–FM modulation vocoder. The TIMIT sentence "George is paranoid about a future gas shortage" is shown in Fig. 6.9 (a). Next, the speech spectrograms are displayed for the original speech signal (b), and for synthetic speech (c)–(e), computed with an analysis window of 15 msecs.

In Fig. 6.9 (c), we display the spectrogram for the output of the analysis–synthesis system of Fig. 6.1, i.e., the amplitude envelope and instantaneous frequency signals are neither modeled nor coded. The differences between (b) and (c) are mainly introduced by the time–varying filterbank, which does not guarantee perfect reconstruction. Most of the differences between the two spectrograms are concentrated in the low frequency band 0–400 Hz, the high frequency band 4–5 kHz and spectral valleys. Perceptually, the most important distortion is the absence of the 0–400 Hz band [31], which we henceforth refer as the *zeroth band*. We have observed that when the zeroth band is added to the AM–FM analysis–synthesis scheme, synthetic speech sounds very close to the original. Overall, the quality and naturalness of the analysis–synthesis system is very good, the main difference from the original speech being the absence of the 0–400 Hz band. We are currently implementing a version of the analysis–synthesis system that explicitly models a fixed–in–frequency zeroth band.[15]

In Fig. 6.9 (d), we display the spectrogram of the speech signal that was reconstructed from the multipulse modeled amplitude envelope and instantaneous frequency resonant signals, as discussed in Section 6.3. The spectrograms in (c) and (d) look very similar. In general, speech reconstructed from the modeled information signals sounds very natural and almost identical to the original. Some differences can be heard for unvoiced fricatives because the multipulse amplitude model represents better the voiced than the unvoiced speech regions. In general, though, the information signal modeling is very accurate.

Finally, in Fig. 6.9 (e), the spectrogram of the speech signal coded at 4.8 kbits/sec is shown. Overall, the coded speech quality is good. The synthetic speech has sometimes a harsh quality due to the inadequate number of bits assigned to coding the excitation

[15] We have found that adding an extra band greatly simplifies the modeling/coding efforts compared to altering the first time–varying filter to include the zeroth band.

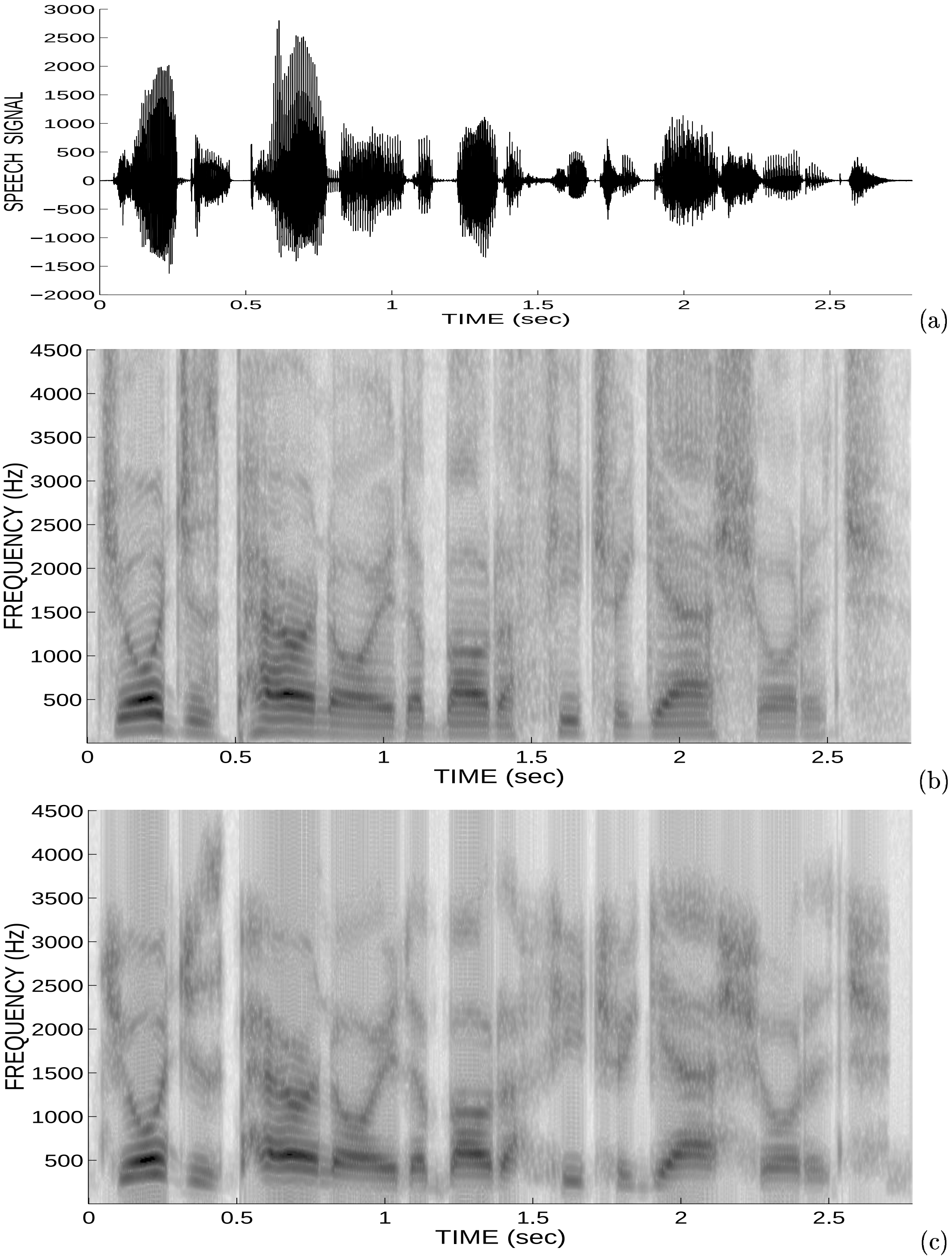


Figure 6.9: (a) Speech signal: "George is paranoid about a future gas shortage". Speech spectrogram: (b) original signal, (c) sum of four formant bands.

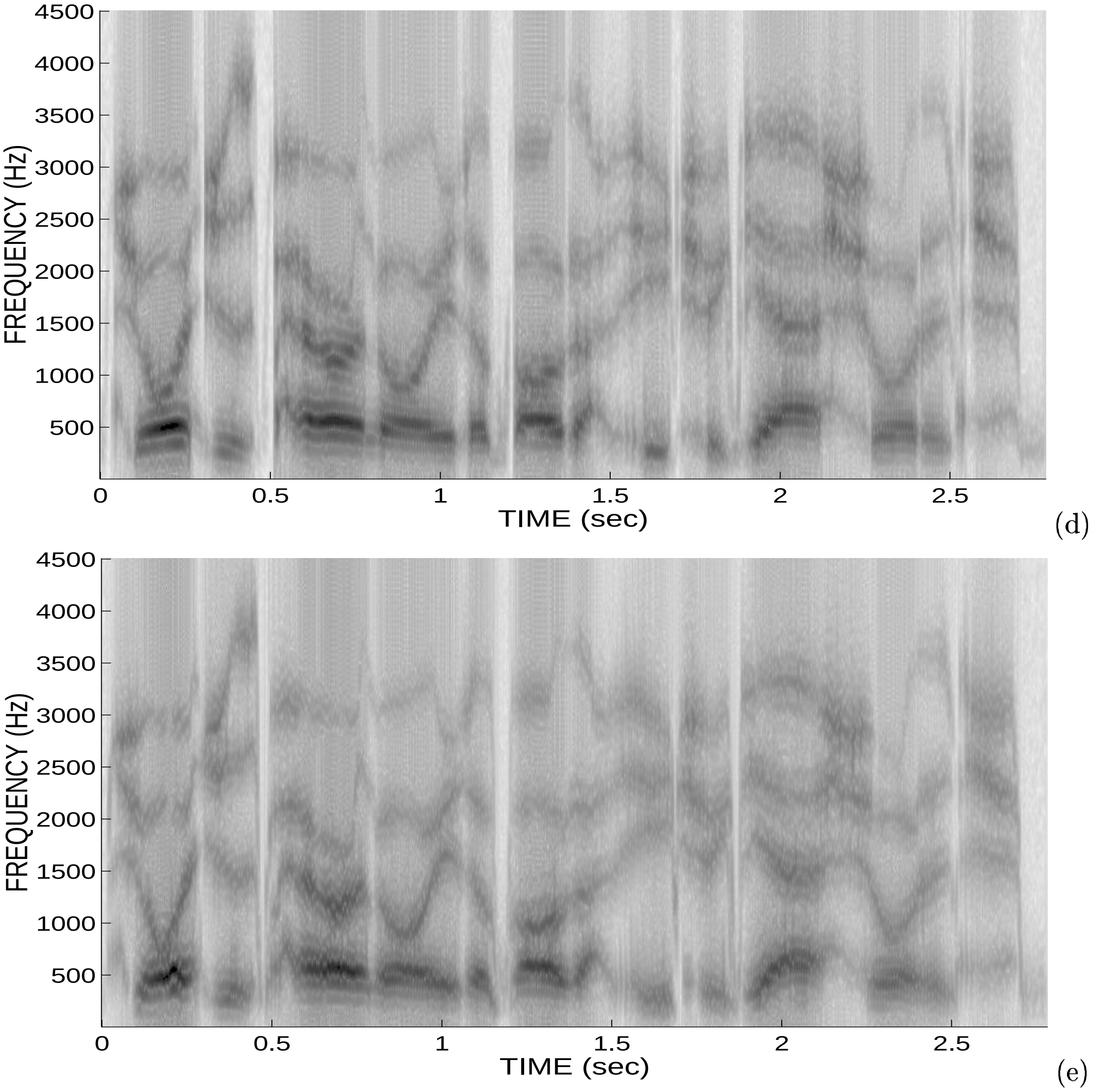


Figure 6.9: (cont.) Speech spectrogram for the sum of four formant bands: reconstructed from the multipulse modeled (d) and coded at 4.8 kbits/sec (e) information signals, respectively.

positions. The lost periodicity is observable also in the spectrogram in (e). The quality improves as the bit rate increases and most of the artifacts disappear at 9.6 kbits/sec. Currently, we are investigating more efficient coding/quantization schemes to get rid of the coding artifacts.

Overall, the AM–FM modulation vocoder is capable of producing speech of natural quality at low bit rates. Comparisons with other speech coders (e.g., multipulse, CELP) are currently being performed.

6.6 Discussion

An important implementation issue for the AM–FM vocoder is the design of the analysis filterbank. Currently, we are using a bank of time–varying Gabor filters. Filters with a flatter passband could improve the quality of the coder. Further, the filterbank can be designed to guarantee perfect reconstruction. Preliminary experiments with FIR filters have shown that when using signals with a flatter passband or larger bandwidth the quality of the synthetic speech of the analysis–synthesis system improves. Note, though, that for increased filter bandwidth, the modeling of the amplitude envelope and instantaneous frequency signals becomes less accurate, due to out–of–band interference (see Section 3.4.3). This tradeoff should be taken into account when designing the analysis filterbank.

More efficient coding/quantization schemes of the amplitude envelope excitation signal can be devised. Specifically, vector quantization can be used to code the excitation pulse amplitudes of different resonant signals. We are currently investigating the issues involved in such a vector quantization scheme.

Two promising applications of the AM–FM modulation model are pitch and time–scale modification. The parametric modeling of the amplitude envelope and instantaneous frequency greatly simplifies these tasks. Note that robust pitch and time–scale modification algorithms are essential for a text–to–speech synthesis application.

Overall, the AM–FM modulation vocoder accounts for a variety of speech production phenomena not described in linear models and, as a result, produces speech of very natural quality. In addition, the detailed parametric modeling of the amplitude envelope and instantaneous frequency signals offers the means to study the perceptual effects of amplitude and frequency modulations in speech resonances. Finally, the AM–FM analysis–synthesis system offers the possibility to modify speech, i.e., altering the speakers characteristic or the speaking style, by changing the amount of amplitude and frequency modulation in different formants. Additional research is needed to quantify how such modifications affect the speech quality.

Chapter 7

Conclusions and Future Research

In this thesis, we have addressed the problem of speech analysis, and coding in the context of a general nonlinear speech model, the AM–FM modulation model. First, a set of speech analysis/demodulation tools were developed and tested. All speech processing applications presented in this dissertation use multiband demodulation as the underlying analysis method. A bank of Gabor filters was used to filter the speech signal and then a demodulation algorithm, e.g., energy separation or Hilbert transform demodulation, was applied on each speech band. The resulting amplitude envelope and instantaneous frequency signals were either used to compute the short–time formant and pitch estimates in speech analysis applications, or modeled and coded in a vocoder application. Overall, the AM–FM modulation model and multiband demodulation analysis are a general nonlinear approach to speech processing with a wide range of successful applications.

In this chapter, we summarize the main contributions of this dissertation. We also propose extensions to the speech analysis and coding work. Finally, two promising speech processing applications for the AM–FM modulation model are explored briefly: text–to–speech synthesis and speaker identification. A starting point is provided for these new research topics. Specifically, we discuss how the modeling and coding schemes for the amplitude envelope and the instantaneous frequency speech resonance signals, proposed in the context of the AM–FM modulation vocoder, can be exploited for synthesis and recognition applications.

7.1 Main Contributions

The main contributions of this research work were in the areas of speech analysis, speech coding and signal demodulation. Next, we discuss the specific contributions in each area.

7.1.1 Speech Analysis

First, a novel method for speech and signal analysis was introduced, namely, multiband demodulation analysis (MDA). The speech signal was filtered through a bank of Gabor bandpass filters and each band was demodulated to amplitude envelope and instantaneous frequency. This analysis method is non–parametric and can capture both time and frequency domain information, e.g., excitation information and formant structure. In addition, multiband filtering followed by the energy separation algorithm has the attractive physical interpretation of tracking the energy of the source that produced the signal and then separating it to the amplitude and the frequency components. Using the amplitude envelope and the instantaneous frequency signals, short–time estimates were proposed for the formant frequency, the formant bandwidth, and the fundamental frequency. The performance of the estimates was evaluated for both synthetic signals and speech. We concluded that the amplitude weighted mean instantaneous frequency F_w and the short–time slope of the phase S_ϕ perform best for formant and pitch estimation, respectively.

Next, formant and pitch estimation algorithms were proposed using the short time estimates F_w and S_ϕ. The algorithms have the advantage of being conceptually simple, non–parametric and easy to implement in parallel. In addition, thanks to the time–domain implementation, increased time resolution for the formant and pitch contours can be obtained at a small additional computational cost. When compared to linear prediction–based algorithms, the formant tracking algorithm was shown to produce smoother and more detailed formant tracks, and more accurate bandwidth estimates. Further, the formant tracker was supplemented with a robust (raw formants to tracks) decision algorithm that works well in the presence of nasalization. Similarly, the pitch estimation algorithm produced very smooth and accurate fundamental frequency contours, and did not suffer from "pitch doubling".

A by–product of the formant and pitch estimation algorithms was the introduction of two useful time–frequency representations, i.e., the mean amplitude weighted instantaneous frequency $F_w(t,\nu)$ (pyknogram), and the short–time phase slope $S_\phi(t,\nu)$. These representations are accurate visualization tools for the formant and harmonic structure of the speech spectra because they are thinned versions of the wideband and narrowband spectrogram, respectively.

7.1.2 Speech Coding

A novel vocoder was introduced which is based on the AM–FM modulation model. Time–varying filters with center frequencies that follow the formant tracks were used to extract three or four formant resonant signals. Each resonant signal was then demodulated to amplitude envelope and instantaneous frequency. Next, efficient algorithms were proposed for modeling and coding of the amplitude envelope and the instantaneous frequency signals. Finally, speech was synthesized as the sum of the reconstructed formant bands. Overall, the AM–FM modulation vocoder is able to describe and model, in detail, nonlinear and time–varying phenomena not accounted for in source–linear-filter based coders.

The proposed algorithm for efficient modeling and coding of the vocoder information signals was a major contribution of this research work. A multipulse excitation model was used for the amplitude envelope signal. The instantaneous frequency signal was modeled as an average formant frequency with frequency modulations around it. In addition, the "fine–scale" phase information at excitation instants was accounted for. Further, the perceptual importance of the modeling parameters was determined for both the amplitude envelope and the instantaneous frequency signals. Finally, coding schemes were proposed for a vocoder operating in the 4.8–9.6 kbits/sec range. A novelty of the coding algorithm was the classification of the amplitude envelope excitation pulses into a "primary" and "secondary" group and, in general, the efficient use of the correlation between information signals of different formants. The detailed modeling produced synthetic speech of very natural quality. Overall, the AM–FM vocoder has the potential of producing excellent speech quality at 2.4–4.8 kbits/sec.

The proposed modeling algorithms were a first step towards a general time–varying AM–FM modulation vocoder where the user has control of the modulation parameters. Using the AM–FM model, synthetic speech of more natural quality can be obtained, e.g., in a text–to–speech synthesis application. Finally, the vocoder is a useful tool in investigating the perceptual importance of nonlinearities in speech.

7.1.3 Signal Demodulation

The energy separation (ESA) and Hilbert transform (HTD) approaches have been compared for AM–FM signal and speech resonance demodulation. From experiments on synthetic AM–FM signals we have found that the HTD approach yields smaller estimation error than the ESA for amplitude/frequency modulation frequencies close to the carrier frequency of the AM–FM signal. As the carrier frequency to information bandwidth ratio increases the ESA performance relative to the HTD improves. In addition, the ESA has smaller computational complexity than the HTD and excellent time resolution. Overall, the algorithms perform similarly for speech resonance demodulation, but the HTD yields smaller errors for formants in the 0–1 kHz range.

Finally, the smooth ESA (SESA) was introduced that uses a 7–point binomial filter to smooth the energy signals. The SESA reduces the mean absolute demodulation error by 20–80% compared to the ESA, depending on the carrier frequency to information signal bandwidth ratio.

7.2 Ongoing Research and Future Directions

Comparisons of the ESA with demodulation schemes other than the Hilbert transform are currently being performed [59]. Such algorithms include the instantaneous exact Prony, linear prediction, and ESA–based instantaneous demodulation schemes [17].

Next, we propose possible extensions to the speech analysis and coding work. Research directions for text–to–speech synthesis, speaker identification, and speech recognition applications are also proposed.

7.2.1 Formant and Pitch Tracking

Next, improvements to the multiband demodulation formant and pitch tracking algorithm are proposed. A global error minimization (raw formants to tracks) decision algorithm could improve the accuracy of the estimated formant tracks. The error can be formulated in a simple two–term form, one term penalizing for deviations from the average formant frequency[1] and the other for discontinuities in the formant tracks. The details of this approach are more cumbersome than it appears at first. Specifically, the possibility of a missing estimate along a formant track must be accounted for and special care must be taken for nasalized speech. Further, efficient pruning of the less probable paths is essential to guarantee reasonable computational complexity. The global error minimization approach is attractive because it is more accurate and simpler than the ad–hoc rule–based decision algorithms. In Section 5.5, we have proposed a similar global error minimization scheme for determining the pitch contour in the multiband demodulation pitch estimation algorithm.

A modified iterative ESA formant tracking algorithm was proposed in Section 4.6. By applying gradient descent on the (frequency) derivative of the formant estimates $\partial F_w(t,\nu)/\partial \nu$, the accuracy of the estimated formant tracks could improve significantly. In this case, the decision algorithms of the iterative ESA and the multiband demodulation formant tracker would be almost identical.

The speech pyknogram is a useful visualization tool because it is a thinned version of the wideband spectrogram. In the pyknogram, the formant tracks are emphasized and readily observable. Modifications to the original formulation of the estimates could improve the clarity of the representation. Post–filtering can also be applied on the pyknogram to achieve this goal. As discussed in Section 4.6, image processing techniques based on mathematical morphology can be used for this task. Similar techniques can be applied to the phase slope time–frequency representation that clearly displays the speech harmonics.

Finally, detailed comparisons with the existing formant and pitch estimation methods should be performed. As a first step, the MDA formant tracker should be formally compared with LPC–based formant trackers. The pitch estimation algorithm should be formally compared with the cepstrum and autocorrelation methods.

[1] A simple way to define the average formant frequency is to set the average F_1, F_2, F_3 to 500, 1500, 2500 Hz, respectively.

7.2.2 Coding

In Sections 6.3 and 6.4, we have discussed how to efficiently model and code the amplitude envelope and the instantaneous frequency signals. In addition, a qualitative description of the perceptual importance of the modeling parameters was given. This knowledge can be used to investigate new (or to modify the proposed) amplitude envelope and instantaneous frequency coding schemes. The end goal of this effort should lead to a low bit rate AM–FM modulation vocoder that operates in the 2.4–4.8 kbits/sec range.

In general, extensive experimentation with a large number of parameters is required to optimize the quantization scheme. The work in this thesis is only a starting point in this direction. We are continuing the effort towards efficient coding and quantization at 4.8 and 9.6 kbits/sec (two implementations).

Another possible improvement to the AM–FM modulation vocoder is to design a better filterbank for the analysis stage. For example, filters with a flatter passband than the Gabor filter can be used. Further, the formant bandwidth information obtained from the formant tracker can be efficiently used in designing the time–varying filterbank.

Finally, formal comparisons with other vocoders with similar bit rate should be performed, e.g., CELP, sinusoidal, and multipulse vocoders.

7.2.3 Speech Recognition

An *energy time–frequency* representation $\mathcal{E}(t, \nu)$ is obtained by filtering the speech signal through a bank of Gabor bandpass filters[2] and applying the energy operator Ψ on the output of each filter. $\mathcal{E}(t, \nu)$ is then defined as the short–time average of the energy around time t and frequency ν. For a fixed time t_0, we define $\mathcal{E}(t_0, \nu)$ to be the *energy spectrum* $\mathcal{ES}(\nu)$ of the corresponding speech frame. The parameters of the energy spectrum are the bandwidth of the Gabor filter and the length of the short–time averaging window. As shown in Fig. 7.1, the energy spectrum can provide a smooth spectral envelope. The energy spectrum yields the mean physical energy required to produce an oscillation, proportional both to amplitude squared and frequency squared. In contrast, the power spectrum yields only the mean square amplitude of an oscillation. Thus, the energy spectrum offers the

[2] The location of the filters ν is fixed in frequency and their spacing follows the Mel–scale [91].

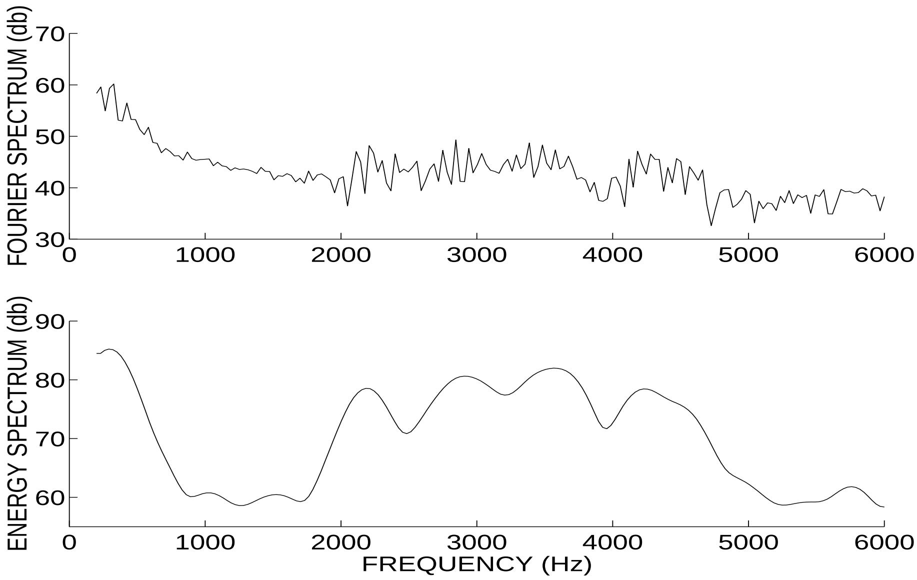


Figure 7.1: The short–time Fourier transform and the energy spectrum for the phoneme /eh/ from the word "zero" sampled at 16 kHz (Gabor bandwidth 280 Hz, 30 msecs window).

means to observe the energy signature (in time) of each formant source and their relation in an analysis frame.

We have used the energy spectrum as the observation vector of an automatic speech recognition system. Recognition was performed using continuous Gaussian density hidden Markov models (HMMs). The energy spectrum was transformed to the cepstrum domain using an inverse cosine transform and the cepstrum coefficients along with their first and second time derivatives where used as the observation vector. The recognition performance was compared with that of a "classic" cepstrum observation vector computed from the power spectrum [91]. Both observation vectors where of the same dimensionality. Despite the differences in the two representations, both sets of features performed almost identically for a phoneme recognition task using 500 sentences of the TIMIT database. Efforts to incorporate other "modulation" features in an HMM speech recognizer had, so far, inconclusive results. New directions is speech recognition using modulation and fine time–scale features are currently being investigated.

7.2.4 Speaker Identification

Modulation and fine time–scale features have been recently used for telephone–based speaker identification [34, 88, 33]. In these experiments, the "modulation features" consisted of a fine pitch estimate computed separately for each formant, and the duration between the primary and a secondary "energy pulse" inside a pitch period. The latter feature was extracted from the maxima of the output of the energy operator applied on the speech resonance signals of a "low" and a "high" formant. These features were used in addition to the the "classic energy,"[3] and the cepstrum coefficients.

Using the modeling ideas introduced in Section 6.3 more robust analysis procedures can be devised to estimate the fine–scale recognition features. Specifically, the analysis–by–synthesis multipulse model can be used to determine both the fine pitch contour, and the duration between primary and secondary pulses in a pitch period. Additional modulation features can be also incorporated in the observation vector. For example, a short–time estimate of the amount of frequency and bandwidth modulation in speech resonances, or the phase information at excitation instances could be used (see Section 6.3).

An important issue is the correlation between the modulation parameters/features and the phonemic content, speaker, or speaking style. This issue can be resolved by using the AM–FM modulation vocoder as the analysis front end to obtain a "modulation" observation vector. Then the statistics of the features can be computed from a large corpus of data labeled by phoneme, speaker, and speaking style, e.g., using the TIMIT and the Lincoln Labs STYLE databases. Depending on the results of these experiments the features should be used for speech, speaker, or speaking style recognition. From preliminary experiments, we have observed that the modulation features are mostly related with the speaker and the speaking style, rather than the phonemic content.

7.2.5 Text–to–Speech Synthesis

One of the most promising applications of the "modulation ideas" is in speech synthesis. We have shown that the detailed modeling of the amplitude envelope and the instantaneous

[3] Again here we refer to the short–time average of the signal squared as the "classic energy" to differentiate from the actual energy of the source that produces the signal.

frequency signals used in the AM–FM modulation vocoder produces speech of very natural quality. AM–FM modeling/coding can be used as the backbone of a text–to–speech system that is either based on concatenation or uses formant transition rules.

Since the AM–FM model is a formant–based model it is expected that an AM–FM text–to–speech system will provide good and natural speech quality during speech transitions. In addition, the modeling of the amplitude and frequency signals greatly simplifies the design of pitch and time–scale modification algorithms, an important part of any text–to–speech system.

The most exciting possibility that the AM–FM modeling offers is the ability to modify the speaking style and/or the voice quality. We have observed for example that by increasing/decreasing the amplitudes of the secondary excitation pulses in the amplitude envelope multipulse model, the quality of the voice ranges from "harsh" to "buzzy". Further, when the amount of frequency modulation in the instantaneous frequency is modified the speech quality is also altered. Additional research is needed to quantify the perceptual effects of such modifications. Finally, it is probable that the amount of amplitude and frequency modulations is correlated with the value of the pitch or the size of the vocal tract. If that is the case, incorporation of the modulation information in the pitch and time–scale modification algorithms could significantly improve speech naturalness, especially for female voices. Overall, the AM–FM modulation model could provide new directions and pave the way to the next generation of text–to–speech systems.

In this thesis, we have attempted to cover a broad area of speech processing applications using the AM–FM modulation model. Further, we have given hints about future research directions in areas not explored in this dissertation, such as text–to–speech synthesis and speaker identification. We sincerely hope that the speech community will continue to show interest and explore the great potential of the "modulation ideas."

Bibliography

[1] M. Abe and S. Ando, "Nonlinear time–frequency domain operators for decomposing sounds into loudness, pitch and timbre," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Detroit, MI), pp. 1368–1372, May 1995.

[2] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, pp. 167–184, Dec. 1982.

[3] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Transactions on Communications*, vol. 30, pp. 600–614, Apr. 1982.

[4] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural–sounding speech at low bit rates," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 614–617, May 1982.

[5] L. Atlas and J. Fang, "Quadratic detectors for general nonlinear analysis of speech," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (San Francisco, CA), pp. II:9–12, Mar. 1992.

[6] B. van der Pol, "The fundamental principles of frequency modulation," *Journal of the IEE (London)*, vol. 93, pp. 153–158, 1946.

[7] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal," *Proceedings of the IEEE*, vol. 80, pp. 520–538, Apr. 1992.

[8] A. C. Bovik, P. Maragos, and T. F. Quatieri, "Measuring amplitude and frequency modulations in noise using multiband energy operators," in *IEEE Internat. Symp. on Time-Frequency and Time-Scale Analysis*, (Victoria, BC, Canada), Oct. 1992.

[9] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM–FM energy detection and separation in noise using multiband energy operators," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3245–3265, Dec. 1993.

[10] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The Dept. of Defense 4.8 kbps standard (proposed federal standard 1016)," in *Advances in Speech Coding* (B. Atal, V. Cuperman, and A. Gersho, eds.), pp. 121–133, Boston, MA: Kluwer Academic Publishers, 1991.

[11] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source–tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 663–671, July 1994.

[12] J. M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the Audio Engineering Society*, vol. 21, July 1973.

[13] J. M. Chowning, "Frequency modulation synthesis of the singing voice," in *Current Directions in Computer Music Research* (M. V. Mathews and J. R. Pierce, eds.), pp. 57–63, Cambridge, MA: MIT Press, 1989.

[14] L. Cohen, "What is a multicomponent signal?," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, pp. V:113–116, 1992.

[15] L. Cohen and C. Lee, "Instantaneous bandwidth," in *Time Frequency Signal Analysis – Methods and Applications* (B. Boashash, ed.), London: Longman–Chesire, 1992.

[16] G. Duncan and M. A. Jack, "Formant estimation algorithm based on pole focusing offering improved noise tolerance and feature resolution," *IEE Proceedings*, vol. 135, Pt. F, pp. 18–32, Feb. 1988.

[17] L. Fertig and J. H. McClellan, "Instantaneous frequency estimation using linear prediction; with comparisons to the DESAs," *IEEE Signal Processing Letters*, submitted, 1994.

[18] J. L. Flanagan, *Speech Analysis Synthesis and Perception.* Berlin: Springer-Verlag, 2nd ed., 1972.

[19] J. L. Flanagan, “Parametric coding of speech spectra,” *Journal of the Acoustical Society of America*, vol. 68, pp. 412–419, Aug. 1980.

[20] J. L. Flanagan and S. W. Christensen, “Computer studies on parametric coding of speech spectra,” *Journal of the Acoustical Society of America*, vol. 68, pp. 420–430, Aug. 1980.

[21] D. H. Friedman, “Instantaneous frequency distribution vs. time: an interpretation of the phase structure of speech,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Tampa, Florida), pp. 1121–1124, Mar. 1985.

[22] D. Gabor, “Theory of communication,” *Journal of the IEE (London)*, vol. 93, pp. 429–457, 1946.

[23] E. B. George, *An Analysis–by–Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Signal Processing.* PhD thesis, Georgia Institute of Technology, Atlanta, GA, Nov. 1991.

[24] B. Gold and L. Rabiner, “Parallel processing techniques for estimating pitch periods of speech in the time domain,” *Journal of the Acoustical Society of America*, vol. 46, no. 2, pp. 442–448, 1969.

[25] D. W. Griffin and J. S. Lim, “The multi–band excitation vocoder,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1223–1235, Aug. 1988.

[26] H. M. Hanson, P. Maragos, and A. Potamianos, “A system for finding speech formants and modulations via energy separation,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 436–443, July 1994.

[27] G. C. Hegerl and H. Hoge, “Numerical simulation of the glottal flow by a model based on the compressible Navier-Stokes equations,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Toronto, Ontario), May 1991.

[28] H. Herzel, I. Steinecke, W. Mende, and K. Wermke, “Chaos and bifurcations during voiced speech,” in *Complexity, Chaos and Biological Evolution* (E. Mosekilde and L. Mosekilde, eds.), pp. 41–50, New York, NY: Plenum Press, 1991.

[29] W. Hess, *Pitch Determination of Speech Signals.* New York, NY: Springer–Verlag, 1983.

[30] J. N. Holmes, "Formant excitation before and after glottal closure," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, Apr. 1976.

[31] J. N. Holmes, "Formant synthesizers: Cascade or parallel?," *Speech Communication*, vol. 2, pp. 251–273, Dec. 1983.

[32] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two–mass model of the vocal cords," *Bell System Technical Journal*, vol. 51, pp. 1233–1268, July 1972.

[33] C. R. Jankowski, T. F. Quatieri, and D. A. Reynolds, "Formant AM–FM for speaker identification," in *Proc. IEEE–SP Internat. Conf. on Time–Frequency and Time–Scale Analysis*, (Philadelphia, PA), Oct. 1994.

[34] C. R. Jankowski, T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Detroit, MI), May 1995.

[35] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice–Hall, 1984.

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

[37] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Albuquerque, New Mexico), pp. 381–384, Apr. 1990.

[38] J. F. Kaiser, "On Teager's energy algorithm and its generalization to continuous signals," in *IEEE DSP Workshop*, (New Paltz, NY), Sept. 1990.

[39] J. F. Kaiser, "Some useful properties of the Teager's energy operators," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. III:149–152, Apr. 1993.

[40] J. M. Kates, "A time-domain digital cochlear model," *IEEE Transactions on Signal Processing*, vol. 39, no. 12, pp. 2573–2592, 1991.

[41] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971–995, Mar. 1980.

[42] D. H. Klatt, "Prediction of perceived phonetic distance from critical–band spectra: A first step," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Paris, France), pp. 1278–1282, 1982.

[43] D. H. Klatt, "Review of text–to–speech conversion for english," *Journal of the Acoustical Society of America*, vol. 82, pp. 737–793, Sept. 1987.

[44] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, Feb. 1990.

[45] H. P. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Detroit, MI), pp. 732–736, May 1995.

[46] T. Koizumi, S. Taniguchi, and S. Hiromitsu, "Glottal source–vocal tract interaction," *Journal of the Acoustical Society of America*, vol. 78, pp. 1541–1547, Nov. 1985.

[47] G. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 709–7297, Aug. 1986.

[48] P. Kroon and B. S. Atal, "Pitch predictors with high temporal resolution," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Albuquerque, New Mexico), pp. 661–664, Apr. 1990.

[49] P. Kroon, E. F. Deprettere, and R. J. Sluyter, “Regular-pulse excitation—a novel approach to effective and efficient multipulse coding of speech,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 1054–1063, Oct. 1986.

[50] Y. Laprie and M. Berger, “A new paradigm for reliable automatic formant tracking,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Adelaide, Australia), pp. II:201–205, Apr. 1994.

[51] D. Malah, “Time–domain algorithms for harmonic bandwidth reduction and time scaling of speech signals,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 121–133, Apr. 1979.

[52] L. Mandel, “Interpretation of the instantaneous frequency,” *American Journal of Physics*, vol. 42, pp. 840–846, Oct. 1974.

[53] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “On separating amplitude from frequency modulations using energy operators,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (San Francisco, CA), Mar. 1992.

[54] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “Energy separation in signal modulations with application to speech analysis,” *IEEE Transactions on Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[55] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “On amplitude and frequency demodulation using energy operators,” *IEEE Transactions on Signal Processing*, vol. 41, pp. 1532–1550, Apr. 1993.

[56] P. Maragos and A. Potamianos, “Higher–order differential energy operators,” *IEEE Signal Processing Letters*, vol. 2, Aug. 1995.

[57] P. Maragos, T. F. Quatieri, and J. F. Kaiser, “Detecting nonlinearities in speech using an energy operator,” in *IEEE DSP Workshop*, (New Paltz, NY), Sept. 1990.

[58] P. Maragos, T. F. Quatieri, and J. F. Kaiser, “Speech nonlinearities, modulations and energy operators,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Toronto, Ontario), 1991.

[59] P. Maragos, B. Santhanam, and A. Potamianos, "Co–channel demodulation and AM–FM feaure extraction using energy operators and energy separation," Tech. Rep. 8th monthly, C.R.A.S.P., May 1995.

[60] P. Maragos and R. W. Schafer, "Morphological systems for multidimensional signal processing," *Proceedings of the IEEE*, vol. 78, no. 4, pp. 690–710, 1990.

[61] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, pp. 367–377, Dec. 1972.

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

[63] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Albuquerque, New Mexico), pp. 249–252, Apr. 1990.

[64] R. J. McAulay and T. F. Quatieri, "Low–rate speech coding based on the sinusoidal model," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), pp. 165–208, New York, NY: Marcel Dekker, Inc., 1992.

[65] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, pp. 135–141, Apr. 1974.

[66] A. V. McCree and T. P. Barnwell, "Implementation and evaluation of a 2400 bps mixed excitation LPC vocoder," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Minneapolis, MN), pp. II:159–162, Apr. 1993.

[67] R. McEachern, "How the ear really works," in *Proc. IEEE Internat. Symp. on Time-Frequency and Time-Scale Analysis*, (Victoria, BC, Canada), pp. 437–440, Oct. 1992.

[68] R. H. McEachern, "Hearing it like it is: Audio signal processing the way the ear does it," *DSP Applications*, pp. 35–47, Feb. 1994.

[69] R. S. McGowan, “An aeroacoustic approach to phonation,” *Journal of the Acoustical Society of America*, vol. 83, pp. 696–704, Feb. 1988.

[70] R. S. McGowan, “The quasisteady approximation in speech production,” *Journal of the Acoustical Society of America*, vol. 93, pp. 3011–3013, Nov. 1993.

[71] Y. Medan, E. Yair, and D. Chazan, “Super resolution pitch determination of speech signals,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 39, pp. 40–48, Jan. 1991.

[72] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations.* New York, NY: Wiley, 1979.

[73] M. Niranjan and I. Cox, “Recursive tracking of formants in speech signals,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Adelaide, Australia), pp. II:205–208, Apr. 1994.

[74] A. M. Noll, “Cepstrum pitch determination,” *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–310, 1966.

[75] A. H. Nuttall, “On the quadrature approximation to the Hilbert transform of modulated signals,” *Proceedings of the IEEE*, vol. 54, pp. 1458–1459, 1966.

[76] A. H. Nuttall, “Complex envelope properties, interpretation, filtering and evaluation,” Tech. Rep. TR 8827, Naval Underwater Systems Center, Feb. 1991.

[77] H. Ohmura, “Fine pitch contour extraction by voice fundamental wave filtering method,” in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Adelaide, Australia), pp. II:189–192, Apr. 1994.

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

[79] P. E. Papamichalis, *Practical Approaches to Speech Coding.* Prentice–Hall, 1987.

[80] A. Papoulis, *The Fourier Transform and Its Applications.* New York, NY: McGraw-Hill, 1962.

[81] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* New York: McGraw-Hill, 1984.

[82] A. D. Pierce, *Acoustics: An Introduction to its Physical Principles and Applications.* McGraw-Hill, 1981.

[83] A. Potamianos and P. Maragos, "Applications of speech processing using an AM–FM modulation model and energy operators," in *Proc. European Signal Processing Conf.*, (Edinburgh, Scotland), pp. III:1669–1672, Sept. 1994.

[84] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, pp. 95–120, May 1994.

[85] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Detroit, MI), May 1995.

[86] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *Journal of the Acoustical Society of America*, submitted, 1994.

[87] T. F. Quatieri, R. B. Dunn, P. Maragos, J. F. Kaiser, and A. C. Bovik, "Detection of transient signals using the energy operator," tech. rep., MIT Lincoln Labs, in preparation.

[88] T. F. Quatieri, C. R. Jankowski, and D. A. Reynolds, "Energy onset times for speaker identification," *IEEE Signal Processing Letters*, vol. 1, pp. 160–162, Nov. 1994.

[89] T. F. Quatieri, J. F. Kaiser, and P. Maragos, "Transient detection in AM–FM background using an energy operator," in *IEEE Underwater Acoustic Signal Processing Workshop*, (Univ. of Rhode Island), Oct. 1991.

[90] T. F. Quatieri and R. J. McAulay, "Shape invariant time–scale and pitch modification of speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 40, pp. 497–510, Mar. 1992.

[91] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice–Hall, 1993.

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

[93] M. A. Ramahlo and R. J. Mammone, "New speech enhancement techniques using the pitch mode modulation model," in *Proc. 36th Symp. on Circuits and Systems*, (Detroit, MI), Aug. 1993.

[94] M. J. Ross, F. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, pp. 353–362, Oct. 1974.

[95] B. Santhanam and P. Maragos, "Demodulation of two–component AM–FM signal mixtures," Tech. Rep. 94–13, DSP Lab, Georgia Tech, Nov. 1994.

[96] M. R. Schroeder and B. S. Atal, "Code–excited linear prediction (CELP): High–quality speech at very low bit rates," in *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing*, (Tampa, Florida), pp. 937–940, Mar. 1985.

[97] M. Schwartz, *Information Transmission, Modulation and Noise.* New York, NY: McGraw-Hill, 1980.

[98] J. Serra, *Image Analysis and Mathematical Morphology.* New York: Academic Press, 1982.

[99] S. Singhal and B. S. Atal, "Amplitude optimization and pitch prediction in multipulse coders," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 317–327, Mar. 1989.

[100] M. M. Sondhi and J. Schroeter, "A hybrid time–frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 955–967, July 1987.

[101] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 599–601, Oct. 1980.

[102] H. M. Teager and S. M. Teager, “A phenomenological model for vowel production in the vocal tract,” in *Speech Sciences: Recent Advances* (R. G. Daniloff, ed.), pp. 73–109, San Diego, CA: College–Hill Press, 1985.

[103] H. M. Teager and S. M. Teager, “Evidence of nonlinear sound production mechanisms in the vocal tract,” in *Speech Production and Speech Modeling* (W. J. Hardcastle and A. Marchal, eds.), pp. 241–261, Boston, MA: Kluwer Academic Publishers, 1990.

[104] T. J. Thomas, “A finite element model of fluid flow in the vocal tract,” *Computer Speech and Language*, vol. 1, pp. 131–151, 1986.

[105] T. Toyoshima, N. Miki, and N. Nagai, “Adaptive formant estimation with compensation for gross spectral shape,” *Electronics and Communications in Japan*, vol. 74–3, pp. 58–67, June 1991.

[106] D. J. Tritton, *Physical Fluid Dynamics.* New York, NY: Oxford University Press, 1988.

[107] C. K. Un, “A low–rate digital formant vocoder,” *IEEE Transactions on Communications*, vol. 26, pp. 344–355, Mar. 1978.

[108] J. van den Berg, J. T. Zantema, and P. Doornenbal Jr., “On the air resistance and the Bernoulli effect of the human larynx,” *Bell System Technical Journal*, vol. 29, pp. 626–631, May 1957.

[109] L. M. van Immerseel and J.-P. Martens, “Pitch and voiced/unvoiced determination with an auditory model,” *Journal of the Acoustical Society of America*, vol. 91, pp. 3511–3526, June 1992.

[110] J. Ville, “Theory et applications de la notion de signal analytique,” *Cable et Transmission*, vol. 2A, pp. 61–74, 1948.

[111] J. D. Wise, J. R. Caprio, and T. W. Parks, “Maximum likelihood pitch estimation,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 418–423, Oct. 1976.

[112] A. Zayezdny and I. Druckmann, "A new method of signal description and its applications to signal processing," *Signal Processing*, vol. 22, pp. 153–178, Feb. 1991.