Speech formant frequency and bandwidth tracking using multiband energy demodulation

Alexandros Potamianos and Petros Maragos
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332-0250, USA

December 6, 1996

To appear in the *Journal of the Acoustical Society of America*

Running title: *Multiband demodulation formant tracking*

Received: _____

Abstract

In this paper, the AM–FM modulation model and a multiband demodulation analysis scheme are applied to formant frequency and bandwidth tracking of speech signals. Filtering by a bank of Gabor bandpass filters is performed to isolate each speech resonance in the signal. Next, the amplitude envelope (AM) and instantaneous frequency (FM) are estimated for each band using the energy separation algorithm (ESA). Short-time formant frequency and bandwidth estimates are obtained from the instantaneous amplitude and frequency signals; two frequency estimates are proposed and their relative merits are discussed. The short-time estimates are used to compute the formant locations and bandwidths. Performance and computational issues of the algorithm are discussed. Overall, multiband demodulation analysis (MDA) is shown to be a useful tool for extracting information from the speech resonances in the time-frequency plane.

PACS number: 43.72.Ar

Introduction

Motivated by several nonlinear and time-varying phenomena during speech production, Maragos, Quatieri, and Kaiser (1991, 1993a) proposed an AM–FM modulation model that represents a single speech resonance $r(t)$ as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure

$$r(t) = a(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \tag{1}$$

where $f_c \stackrel{\triangle}{=} F$ 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, the speech signal $s(t)$ is modeled as the sum $s(t) = \sum_{k=1}^{N} r_k(t)$ of N such AM–FM signals, one for each formant.

To obtain the amplitude envelope $|a(t)|$ and the instantaneous frequency $f(t)$ signals from a speech resonance $r(t)$, a demodulation algorithm must be used. In addition, a filtering scheme is needed to isolate a single resonance signal $r(t)$ from the speech signal before demodulation can be performed. These two steps of speech analysis in the framework of the AM–FM modulation model were systematically introduced by Bovik, Maragos and Quatieri (1993) and will be henceforth

referred to as *multiband demodulation analysis* (MDA). The representation of a speech signal $s(t)$ by the formant amplitude envelope and instantaneous frequency signals is rich because it reveals both the spectral structure and the excitation timing information of different formant bands. The modulation model can also account for nonlinear phenomena during speech production, e.g., energy transfer among excitation source(s) and resonators in the vocal tract. Teager and Teager (1990) presented experimental evidence of vorticity and unstable separated airflow during vowel production; as a result the energy and frequency of speech resonances may vary with time (Maragos *et al.* 1993a). Further, Ananthapadmanabha and Fant (1982) have shown that source–vocal tract interaction gives rise to a frequency modulation component in the resonant frequencies, i.e., $f(t)$ is time-varying within a pitch period. Finally, as the vocal tract shape changes during phonemic transitions, flow instabilities can arise (Tritton 1988). The AM–FM modulation model can analyze such phenomena (indirectly) by measuring the modulations present at each speech resonance.

Formant tracking is an important speech analysis problem since formant location is a very important cue for human and machine speech recognition. In addition, formant trajectories have been used successfully in both speech coding and speech synthesis applications. Most formant tracking algorithms are based on linear prediction (LP) analysis (McCandless 1974; Duncan and Jack 1988) and 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 the deficiencies of LP by using a pole-zero model for formant tracking (Toyoshima *et al.* 1991). Other more complex formant tracking algorithms use the extended Kalman filter (Niranjan and Cox 1994) or hidden Markov models (Kopec 1986). Alternatively, we propose here a multiband demodulation approach to formant tracking in the framework of the AM–FM modulation model that is easy to implement and overcomes most of the deficiencies of LP.

In this paper, 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 band: the *mean instantaneous frequency*, which has been used for formant tracking by Hanson *et al.* (1994) and the *mean amplitude weighted instantaneous frequency*, a time domain equivalent of the first central spectral moment (Cohen and Lee 1992). Based on the weighted frequency estimate, the modulation model, and a multiband filtering demodulation scheme, we propose the *multiband demodulation formant tracker*. The algorithm produces reliable formant tracks and realistic formant bandwidth

estimates. In addition, it is simple, easy to implement, and avoids most of the drawbacks of LP–based formant trackers.

The organization of the paper is as follows. First, the analysis tools of the modulation model are presented, i.e., multiband filtering and demodulation. In Section II, the unweighted and weighted short-time average frequency estimates are proposed. The performance of the formant frequency estimates is evaluated for both synthetic and real speech signals. The multiband formant tracking algorithm is introduced in Section III. The speech signal is analyzed through a bank of Gabor filters, each band is demodulated, and the formant frequency and bandwidth estimates are computed for each band. Next, a decision algorithm is presented that converts the short-time estimates to raw formants and, ultimately, to formant tracks. Finally, in Sections IV and V performance and implementation issues are discussed.

I Multiband Filtering and Demodulation

A speech resonance is extracted from the speech signal through filtering. A real Gabor bandpass filter is used for this purpose with impulse response $h(t)$ and frequency response $H(f)$

$$h(t) = \exp(-\alpha^2 t^2)\cos(2\pi\nu t) \tag{2}$$

$$H(f) = \frac{\sqrt{\pi}}{2\alpha}\left(\exp\left[-\frac{\pi^2(f-\nu)^2}{\alpha^2}\right] + \exp\left[-\frac{\pi^2(f+\nu)^2}{\alpha^2}\right]\right) \tag{3}$$

where ν is the center frequency of the filter chosen equal to the formant frequency F, and α is the bandwidth parameter. The effective RMS bandwidth of the filter was defined by Gabor (1946) as $\sqrt{2\pi}$ times the RMS bandwidth, and is equal to $\alpha/\sqrt{2\pi}$. In discrete time, the impulse response is a sampled and truncated version of Eq. (2).

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 domains provides accurate amplitude and frequency estimates in the demodulation stage that follows. In Bovik *et al.* (1993), one can find a detailed discussion on the advantages of Gabor wavelets for multiband energy demodulation.

The *energy separation algorithm* (ESA) was developed by Maragos, Kaiser and Quatieri (1993a) to demodulate a speech resonance $r(t)$ into amplitude envelope $|a(t)|$ and instantaneous frequency $f(t)$ signals. The ESA is based on an energy-tracking operator invented by Teager and systematically introduced by Kaiser (1990). The energy operator tracks the energy of the source producing

an oscillation signal $s(t)$ and is defined as

$$\Psi[s(t)] = [\dot{s}(t)]^2 - s(t)\ddot{s}(t) \tag{4}$$

where $\dot{s} = ds/dt$. The ESA frequency and amplitude estimates are

$$\frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{s}(t)]}{\Psi[s(t)]}} \approx f(t)\ , \qquad \frac{\Psi[s(t)]}{\sqrt{\Psi[\dot{s}(t)]}} \approx |a(t)|. \tag{5}$$

Similar equations and algorithms exist in discrete time (Maragos *et al.* 1993a, 1993b). The ESA is simple, computationally efficient, and has excellent time resolution.

An alternative way to estimate $|a(t)|$ and $f(t)$ is through the Hilbert transform demodulation (HTD), i.e., as the modulus and the phase derivative of the Gabor analytic signal (Papoulis 1984). In Potamianos and Maragos (1994b), it is shown that the HTD and the ESA produce similar results for speech resonance demodulation, but the HTD has higher computational complexity. Further, 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 paper.

II 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 $f(t)$

$$F_u = \tfrac{1}{T} \int_{t_0}^{t_0+T} f(t)\, dt \tag{6}$$

$$[B_u]^2 = \tfrac{1}{T} \int_{t_0}^{t_0+T} (f(t) - F_u)^2\, dt \tag{7}$$

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

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)\,[a(t)]^2\, dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{8}$$

$$[B_w]^2 = \frac{\int_{t_0}^{t_0+T} [(\dot{a}(t)/2\pi)^2 + (f(t) - F_w)^2 [a(t)]^2] dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{9}$$

where the additional term $(\dot{a}(t)/2\pi)^2$ in B_w accounts for the amplitude modulation contribution to the bandwidth (Cohen and Lee 1992)

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

$$x(t) = a_1(t)\cos[2\pi f_1 t] + a_2(t)\cos[2\pi f_2 t] \qquad t \in [0, 0.1] \text{ sec} \tag{10}$$

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. 1 (a)-(d) we display the amplitude envelope $|a(t)|$ and the instantaneous frequency $f(t)$ 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 amplitude envelope contours shown in (a), (b). At envelope maxima the instantaneous frequency computed via the HTD (shown in (c)) is equal to the average (amplitude weighted) frequency of the two sinusoids $f = (a_1 f_1 + a_2 f_2)/(a_1 + a_2)$, while at envelope minima f presents spikes of value $f = (a_1 f_1 - a_2 f_2)/(a_1 - a_2)$, i.e., the spikes point towards the frequency of the sinusoid with the larger amplitude (see Appendix A). The ESA and HTD frequency estimates take similar values, yet the orientation of the instantaneous frequency spikes in (c), (d) is somewhat different. As discussed in the appendix, the spikes in the ESA estimate of f point toward the frequency of the sinusoid with the larger amplitude frequency product (the turning point in (d) is where the dotted lines cross), 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. 1(e); F_u locks onto the sinusoid with the greater amplitude (amplitude frequency product for the ESA). The weighted estimate F_w, shown in (f), 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 explanation in the appendix).

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 (McAulay and Quatieri 1986); in particular, a speech resonance can be modeled as a sum of a few sinusoids. The behavior of the F_u, F_w estimates for a speech formant can then be viewed as a generalization of the two sinusoids case analyzed above. For a speech resonance signal, F_u has the tendency to lock on the frequency with the greatest amplitude in the formant band, while F_w weights each frequency in the formant band with its

squared amplitude.

In Fig. 2(a), we show the Fourier spectrum of a 25 msecs speech segment and the frequency response of the Gabor filter centered at $\nu = F = 1600$ Hz with effective RMS bandwidth of 440 Hz. The Fourier spectrum of the formant band (Gabor filtered 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. 2(c) and (d) we use a Gabor filter that is centered at 1300 Hz, 300 Hz off the formant frequency. F_u still locks on the harmonic with the greatest amplitude in the spectrum, which is the major formant harmonic. The weighted estimate F_w, being an "average" frequency, deviates from the formant frequency by almost 200 Hz. In this case, the spikes of the instantaneous frequency point towards the formant and the unweighted estimate F_u is a better formant estimate than F_w. There are cases, though, where a single prominent harmonic does not exist "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 Hanson *et al.* (1994) and Section IV). The weighted frequency estimate F_w provides more accurate formant frequencies and is more robust for low energy or noisy frequency bands.

Similarly, the B_w bandwidth estimates is more robust than the B_u estimate. For example, in Fig. 1(d), (e) we display B_u and B_w (computed via the HTD) for the sum of two sinusoids of Eq. (10). 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. As noted below, B_w is the RMS formant bandwidth. Henceforth, B_w is used as the formant bandwidth estimate.

The (squared amplitude) weighted estimates F_w and B_w are time domain equivalents of the first and second central spectral moments of the signal (Ville 1948; Mandel 1974; Cohen and Lee 1992; Potamianos and Maragos 1994b). 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 V). Note that since B_w equals the second spectral moment, B_w is by definition the RMS bandwidth of the 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 (Potamianos and Maragos 1994b). According to the ESA error bounds formulated by Maragos *et al.* (1993a) the

performance of the ESA deteriorates as the carrier frequency (formant) approaches the modulation frequency (fundamental). Thus for frequency bands centered close to the fundamental frequency the HTD can produce smoother estimates than the ESA, when a careful and computationally expensive implementation is used for the discrete-time HTD. In practice, for frequency bands in the 0-500 Hz range, the short-time frequency and (especially) bandwidth estimates B_w are more accurate when computed by the HTD than the ESA. If accurate formant bandwidth estimates are needed in this low frequency range the HTD should be used for demodulation; otherwise the ESA should be used for computational efficiency.

III 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 $|a(t)|$ and instantaneous frequency $f(t)$ 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. (8), (9)) 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. 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. 3(c) represents the range of F_w. In (c), the formants 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. 3(b). We refer to the time-frequency representation of Fig. 3(c) as the *speech pyknogram* ("pyknogram" stems from the Greek word "pykno" ($\pi\upsilon\kappa\nu\acute{o}\varsigma$) = dense). The pyknogram displays clearly the formant positions (and bandwidths) and possibly the location of the spectral zeros (low density areas). Note that a similar time-frequency representation has been proposed by Friedman (1985), where for each frequency band the instantaneous frequency signal is computed, smoothed in the frequency and time domains and displayed vs. time.

In Fig. 4, we show the frequency $F_w(\nu,t_0)$ and bandwidth $B_w(\nu,t_0)$ estimates vs. the center frequency of the Gabor filters ν, for a single analysis frame centered at t_0. 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 frequencies where the Gabor filter center frequency ν and the short-time frequency estimate $F_w(\nu)$ are equal, i.e., $\{\nu : F_w(\nu) = \nu\}$. Yet, we have observed from synthetic and real speech experiments that for a "weak" formant the $\{\nu : F_w(\nu) = \nu\}$ estimate is biased towards the frequency of a neighboring "strong" formant. As a result the second and higher formant tracks 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 because 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: $\{\nu : F_w(\nu) = \nu\}$ 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, $\partial F_w(\nu)/\partial\nu|_{\nu_0}$ 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 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. In this paper, 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 a speech analysis frame centered at time t the raw formants RF are obtained from the time-frequency distribution $F_w(t, \nu)$ as follows:

$$RF_1 = \{\nu : (F_w(\nu) = \nu) \text{ and } (\frac{\partial F_w(\nu)}{\partial\nu} < 0.7) \text{ and } (\nu < 500)\} \quad (11)$$

$$RF_2 = \{F_w(\nu) : (\frac{\partial^2 F_w(\nu)}{\partial\nu^2} = 0) \text{ and } (\frac{\partial F_w(\nu)}{\partial\nu} < 0.7) \text{ and } (\nu > 500)\} \quad (12)$$

$$RF = RF_1 \bigcup RF_2 \quad (13)$$

where $\bigcup$ denotes set union.

In Fig. 5(a), we display the raw formant estimates for the sentence of Fig. 3(a). A three-point binomial smoother is applied on $F_w(t,\nu)$ in the time domain before the raw formant estimates are computed. In Fig. 5(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 (McCandless 1974). 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". The decision algorithm consists of three steps. First, we search for anchor formant segments, i.e., segments where the formants 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. Alternative formant decision algorithms based on evaluating all possible combinations of raw formants to formant tracks can be found in the literature, e.g, hidden Markov model decoding (Kopec 1986) or a functional minimization approach (Laprie and Berger 1994).

IV 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. 6. Speech was synthesized using the tracks shown as dotted lines in Fig. 6(b). 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. 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. 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. 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 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 and 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. 7, we display the LP raw formant frequency and bandwidth estimates for comparison with the MDA estimates in Fig. 5. Although the long-term formant trajectory shapes look similar (except for nasalized speech, where the MDA formant tracker sometimes produces an extra low-frequency formant) there are some important differences over small scales. LP produces a number

of spurious formants that may confuse the formant decision algorithm. 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 (by tracking an extra "nasal formant"), provides good formant bandwidth estimates, and produces very few spurious raw formants. Currently, the MDA formant tracker is being integrated into the *AM–FM modulation vocoder* (Potamianos and Maragos 1994a).

An iterative demodulation algorithm for formant tracking has been proposed by Hanson, Maragos and Potamianos (1994). 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 average 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 iteration to iteration. 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 than the iterative ESA especially in 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.

V 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. Increasing the spacing of the bandpass filters with frequency, decreases the frequency discrimination for higher formants. This is compatible with the formant frequency perceptual resolution (limens) of the ear. In Hanson *et al.* (1994), the performance of the iterative ESA formant tracker has improved by using constant-Q filters.

As discussed in Section II, the choice of the unweighted F_u vs. the weighted F_w frequency estimates is 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 II, that the F_w and B_w estimates are equivalent to the the first and second spectral moments computed in the frequency domain via 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 are used. A 1024-point FFT gives good results for sampling frequency at 16 kHz and a short-time 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 time resolution of the formant tracks at a small computational cost by simply decreasing the size of the short-time averaging window in a second pass of the algorithm.

Next we propose an alternative formant decision algorithm that applies 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. 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; it has found many applications in image processing and nonlinear filtering (Serra 1992; Maragos and Schafer 1990). As a continuation of the work in this paper, we plan to apply algorithms from morphological image analysis for cleaning, segmentation, and thining of the formant tracks in the pyknogram.

Finally, one could possibly use multiband demodulation for spectral zero tracking. In Fig. 3(c), zeros sometimes manifest themselves as areas of low plot density (e.g., for nasalized sounds an anti-formant can be observed between the second and the third formant track). More work is underway for anti-formant tracking using the multiband demodulation analysis (MDA).

VI Conclusions

In this paper, we have presented a collection of ideas and algorithms for estimating the speech formant parameters and for tracking their evolution in time. The formant tracking algorithm was presented in the the framework of the AM–FM speech modulation model and the main speech analysis tool used was multiband filtering followed by demodulation (MDA). We have shown 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.

Acknowledgments

This work was supported by the US National Science Foundation under Grant MIP-9396301.

Appendix A

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

$$x(t) = \sum_n a_n(t) \cos[2\pi f_n t + \theta_n]. \tag{A1}$$

Assuming that the bandwidth of $x(t)$ is much smaller than the mean carrier frequency (mean of f_n), the quadrature error will be small (Nuttall 1991) and the Gabor analytic signal $z(t)$ of $x(t)$ is

$$z(t) \approx \sum_n a_n(t)\ \exp[j\,(2\pi f_n t + \theta_n)]. \tag{A2}$$

The HTD estimates for the amplitude envelope $|a(t)|$ and instantaneous frequency $f(t)$ are (assuming that $a_n(t)$ is slowly varying compared to $\cos[2\pi f_n t]$)

$$|a(t)| = |z(t)| \approx (\ \textstyle\sum_{n,k}\ a_n(t)\ a_k(t)\ \cos[2\pi(f_n - f_k)t + (\theta_n - \theta_k)]\ )^{\frac{1}{2}} \tag{A3}$$

$$f(t) = \tfrac{d}{dt}\ \angle z(t) \approx (\ \textstyle\sum_{n,k}\ f_n\ a_n(t)\ a_k(t)\ \cos[2\pi(f_n - f_k)t + (\theta_n - \theta_k)]\ )\ /\ [a(t)]^2. \tag{A4}$$

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

$$|a(t)| = (\ a_1^2 + a_2^2 + 2\,a_1\,a_2\ \cos[\Delta\omega\, t]\ )^{\frac{1}{2}} \tag{A5}$$

$$f(t) = (\ a_1^2\, f_1 + a_2^2\, f_2 + a_1\, a_2\, (f_1 + f_2)\, \cos[\Delta\omega\, t]\ )\ /\ [a(t)]^2 \tag{A6}$$

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

$$|a| = |a_1 \pm a_2|\ , \qquad\qquad f = \frac{a_1\, f_1 \pm a_2\, f_2}{a_1 \pm a_2}\ . \tag{A7}$$

Thus, at envelope minima f presents spikes pointing towards the frequency of the sinusoid with the larger amplitude a_n. From Eqs. (A5) and (A6) the short-time frequency estimates F_u and F_w defined in Eqs. (8) and (9) are approximately equal to (depending on the analysis frame boundaries)

$$F_u \approx \begin{cases} f_1, & a_1 > a_2 \\ f_2, & a_1 < a_2 \end{cases} \qquad\qquad F_w \approx \frac{a_1^2\, f_1 + a_2^2\, f_2}{a_1^2 + a_2^2} \tag{A8}$$

i.e., F_u locks onto the frequency component with the larger amplitude, while F_w provides a (squared amplitude) weighted mean frequency.

One can obtain equations similar to (A3), (A4) for the ESA but they are of little intuitive value. Instead we investigate the case of the sum of two amplitude modulated sinusoids. The value of the amplitude envelope $|a|$ and instantaneous frequency f at envelope maxima and minima (derived from the continuous time ESA) are

$$|a| = |a_1 \pm a_2|\ , \qquad\qquad f = \left|\frac{a_1 f_1^2 \pm a_2 f_2^2}{a_1 \pm a_2}\right|^{\frac{1}{2}}\ . \tag{A9}$$

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 $a_n f_n$, while F_w takes values similar to Eq. (A8).

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. (A9) compared to Eq. (A7). In general, the performance of the ESA and the HTD is almost identical for speech formant demodulation provided that the fundamental frequency is much smaller than the formant frequency. For a large-bandwidth multi-component signal, though, the two demodulation algorithms can produce quite different results. There, the ESA frequency estimates are biased (the ESA overestimates the frequencies as can be seen from comparing Eqs. (A7) and (A9)).

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. (A3), (A4)), i.e., F_w weights each frequency with the squared amplitude. The behavior of F_u is more complicated. In general, if the signal consists of 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.

References

Ananthapadmanabha, T. V. and Fant, G. (**1982**). "Calculation of True Glottal Flow and its Components", Speech Communication, **1**, 167–184.

Boashash, B. (**1992**). "Estimating and Interpreting the Instantaneous Frequency of a Signal", Proceedings of the IEEE, **80**, 520–538.

Bovik, A. C., Maragos, P., and Quatieri, T. F. (**1993**). "AM–FM Energy Detection and Separation in Noise Using Multiband Energy Operators", IEEE Transactions on Signal Processing, **41**, 3245–3265.

Cohen, L. and Lee, C. (**1992**), "Instantaneous Bandwidth", in *Time Frequency Signal Analysis – Methods and Applications*, edited by B. Boashash, (Longman–Chesire, London).

Duncan, G. and Jack, M. A. (**1988**). "Formant Estimation Algorithm based on Pole Focusing offering improved Noise Tolerance and Feature Resolution", IEE Proceedings, **135, Pt. F**, 18–32.

Friedman, D. H. (**1985**), Instantaneous Frequency Distribution vs. Time: an Interpretation of the Phase Structure of Speech, in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, 1121–1124.

Gabor, D. (**1946**). "Theory of Communication", Journal of the IEE (London), **93**, 429–457.

Hanson, H. M., Maragos, P., and Potamianos, A. (**1994**). "A System for Finding Speech Formants and Modulations via Energy Separation", IEEE Transactions on Speech and Audio Processing, **2**, 436–443.

Kaiser, J. F. (**1990**), On Teager's Energy Algorithm and Its Generalization to Continuous Signals, in *IEEE DSP Workshop*.

Kopec, G. (**1986**). "Formant Tracking Using Hidden Markov Models and Vector Quantization", IEEE Transactions on Acoustics, Speech and Signal Processing, **34**, 709–7297.

Laprie, Y. and Berger, M. (**1994**), A New Paradigm for Reliable Automatic Formant Tracking, in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, II: 201–205.

Mandel, L. (**1974**). "Interpretation of the Instantaneous Frequency", American Journal of Physics, **42**, 840–846.

Maragos, P., Kaiser, J. F., and Quatieri, T. F. (**1993a**). "Energy Separation in Signal Modulations with Application to Speech Analysis", IEEE Transactions on Signal Processing, **41**, 3024–3051.

Maragos, P., Kaiser, J. F., and Quatieri, T. F. (**1993b**). "On Amplitude and Frequency Demodulation Using Energy Operators", IEEE Transactions on Signal Processing, **41**, 1532–1550.

Maragos, P., Quatieri, T. F., and Kaiser, J. F. (**1991**), Speech Nonlinearities, Modulations and Energy Operators, in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process*, 421–424.

Maragos, P. and Schafer, R. W. (**1990**). "Morphological Systems for Multidimensional Signal Processing", Proceedings of the IEEE, **78**, 690–710.

McAulay, R. J. and Quatieri, T. F. (**1986**). "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Transactions on Acoustics, Speech and Signal Processing, **34**, 744–754.

McCandless, S. S. (**1974**). "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra", IEEE Transactions on Acoustics, Speech and Signal Processing, **22**, 135–141.

Niranjan, M. and Cox, I. (**1994**), Recursive Tracking of Formants in Speech Signals, in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, II: 205–208.

Nuttall, A. H. (**1991**). Complex Envelope Properties, Interpretation, Filtering and Evaluation. Technical Report TR 8827, Naval Underwater Systems Center.

Papoulis, A. (**1984**). *Probability, Random Variables and Stochastic Processes*, (McGraw-Hill, New York).

Potamianos, A. and Maragos, P. (**1994a**), Applications of Speech Processing Using an AM–FM Modulation Model and Energy Operators, in *Proc. European Signal Process. Conf.*, III: 1669–1672.

Potamianos, A. and Maragos, P. (**1994b**). "A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation", Signal Processing, **37**, 95–120.

Serra, J. (**1982**). *Image Analysis and Mathematical Morphology*, (Academic Press, New York).

Teager, H. M. and Teager, S. M. (**1990**), "Evidence of Nonlinear Sound Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and Marchal, A., (Kluwer Academic Publishers, Boston, MA), pp. 241–261.

Toyoshima, T., Miki, N., and Nagai, N. (**1991**). "Adaptive Formant Estimation with Compensation for Gross Spectral Shape", Electronics and Communications in Japan, **74–3**, 58–67.

Tritton, D. J. (**1988**). *Physical Fluid Dynamics*, (Oxford University Press, New York, NY).

Ville, J. (**1948**). "Theory et applications de la notion de signal analytique", Cable et Transmission, **2A**, 61–74.

List of Figures

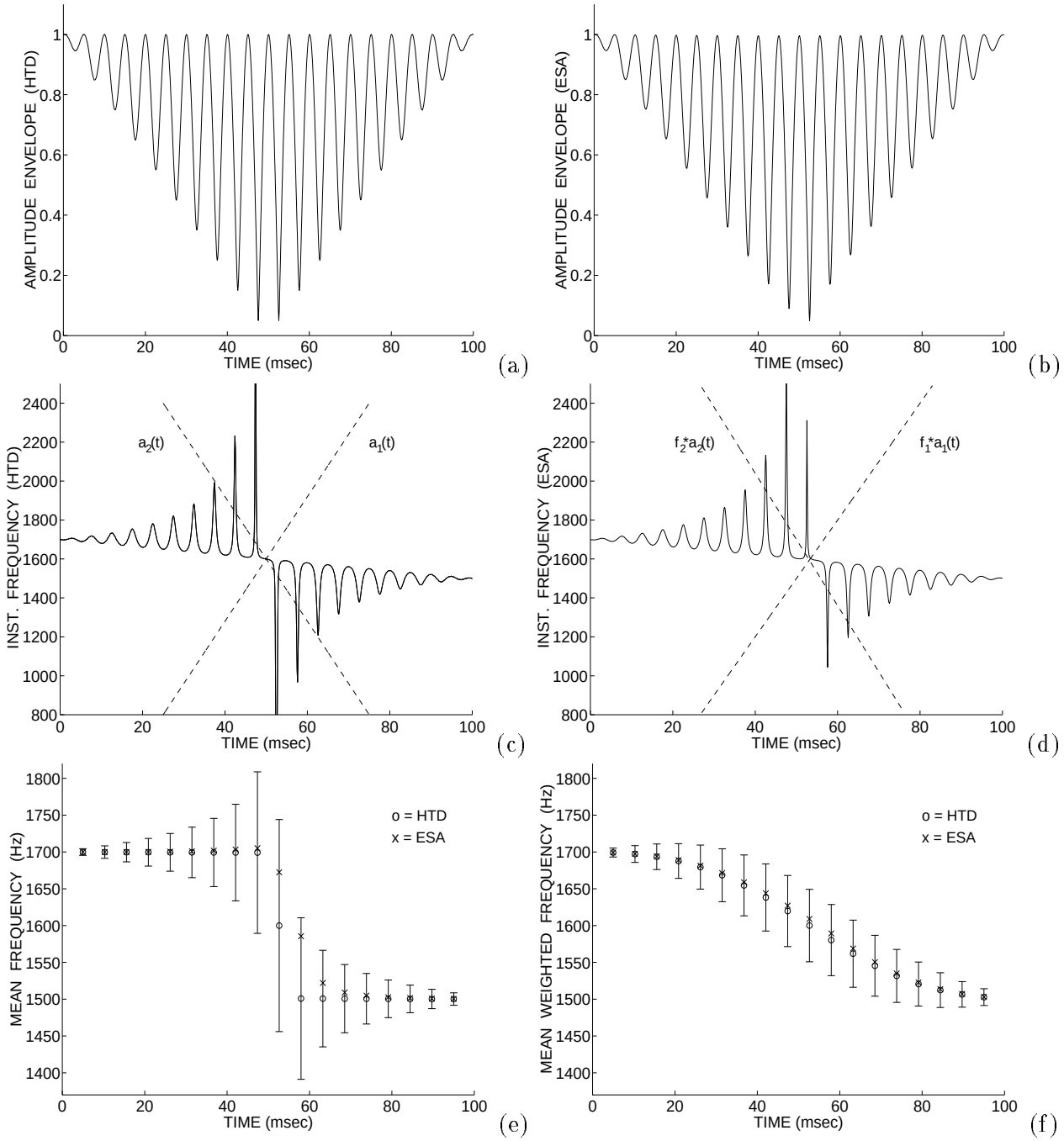

Figure 1: Amplitude envelope and instantaneous frequency of $x(t) = a_1(t)\cos[2\pi f_1 t] + a_2(t)\cos[2\pi f_2 t]$, $a_1(t) = 10\,t$, $a_2(t) = 1 - 10\,t$, $t \in [0, 0.1]$ (sampled at 10 kHz), estimated via HTD (a), (c) and ESA (b), (d). Dotted lines in (c) are proportional to the amplitude of the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

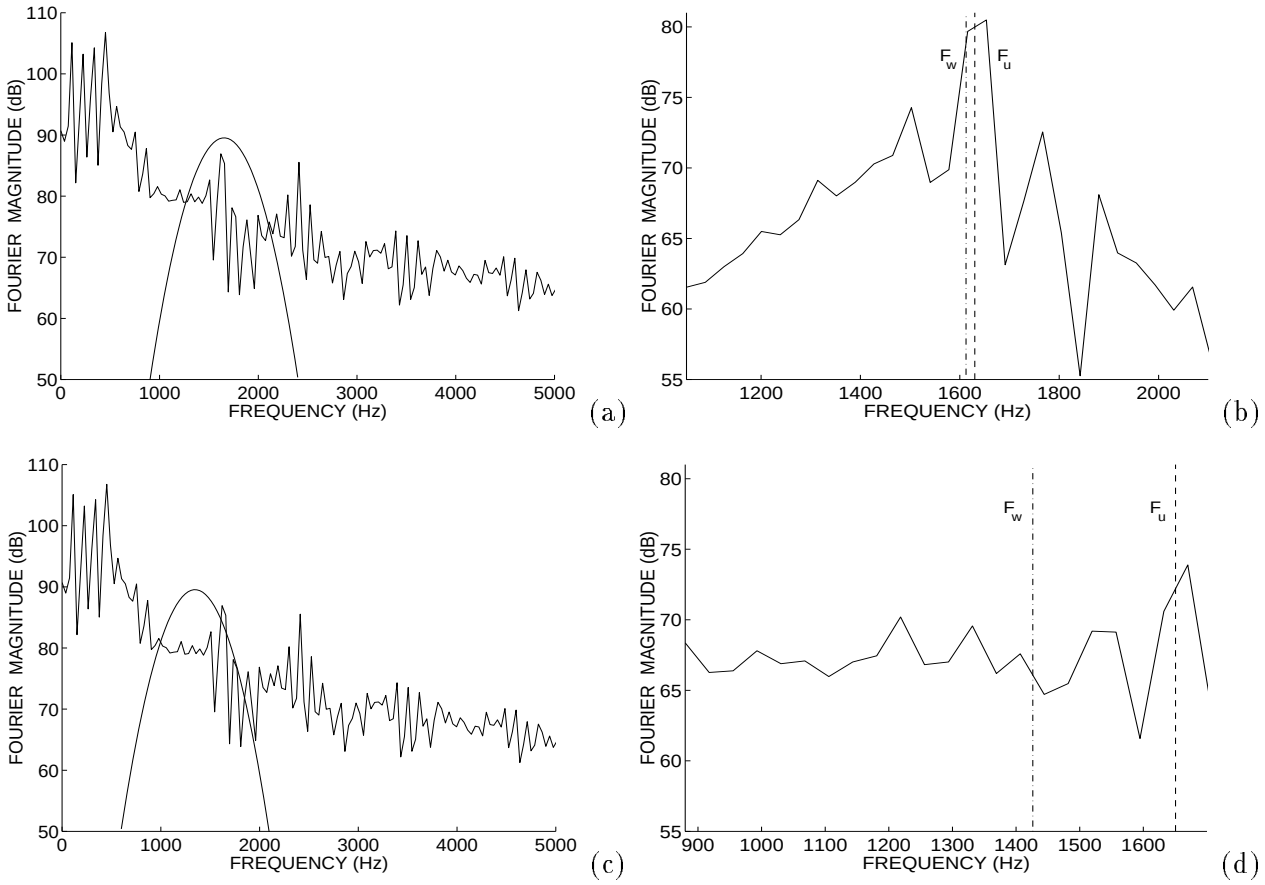

Figure 2: (a) The Fourier spectrum of a 25 msecs segment of speech and the frequency response of the Gabor filter centered at 1600 Hz, (b) the Fourier spectrum of the Gabor bandpass filtered speech (around 1600 Hz); the F_u (dashed line) and F_w estimates (dashed-dotted line). (c),(d) same as (a),(b) but now the Gabor filter is centered at $\nu = 1300$ Hz.

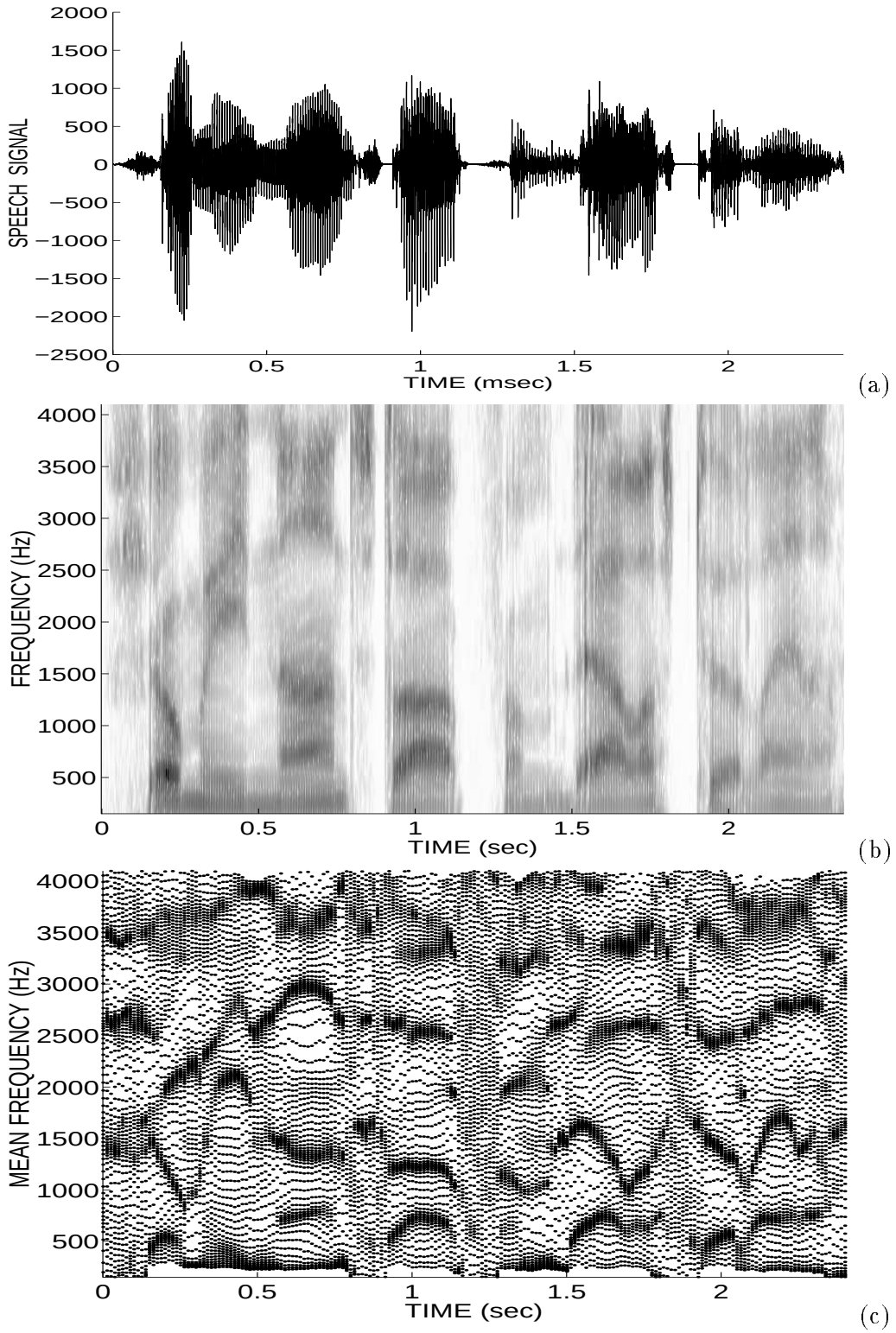

Figure 3: (a) Speech signal: "Show me non-stop from Dallas to Atlanta," (b) wideband spectrogram and (c) pyknogram, i.e., the short-time frequency estimates $F_w(t,\nu)$ for the output of 80 Gabor filters spanning $\nu = 200$ to 4200 Hz displayed vs. time (analysis frame update is 12.5 msecs).

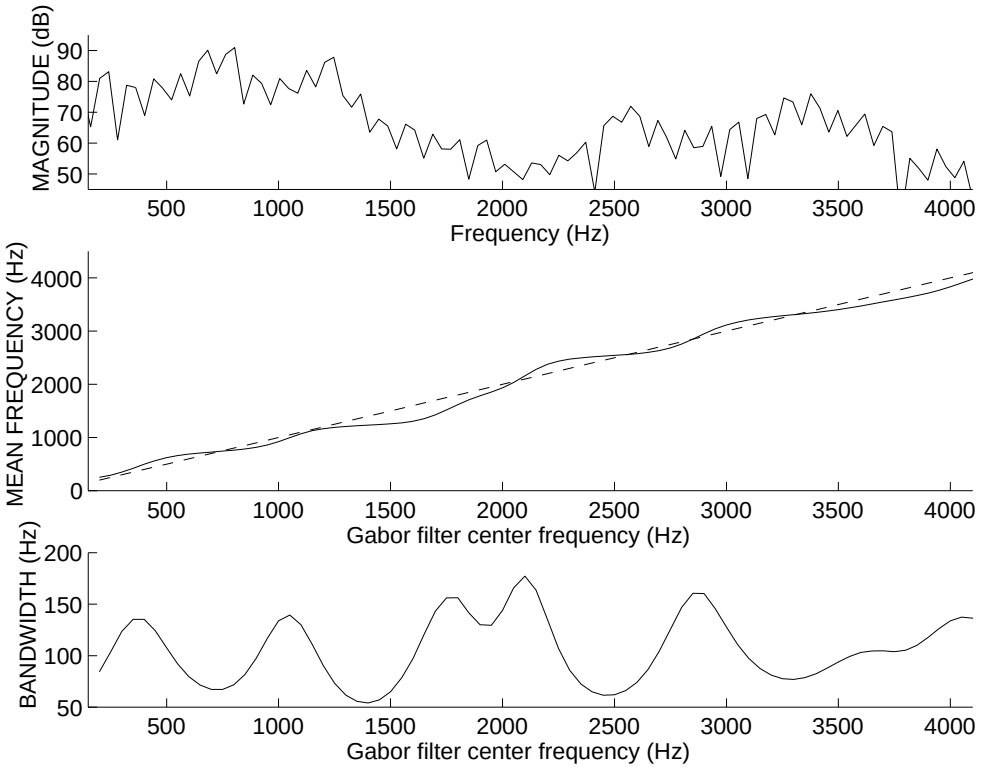

Figure 4: The short-time Fourier transform, the frequency $F_w(f)$ and bandwidth $B_w(f)$ estimates vs. the center frequencies f of the Gabor filters, for a 25 msecs segment of speech.

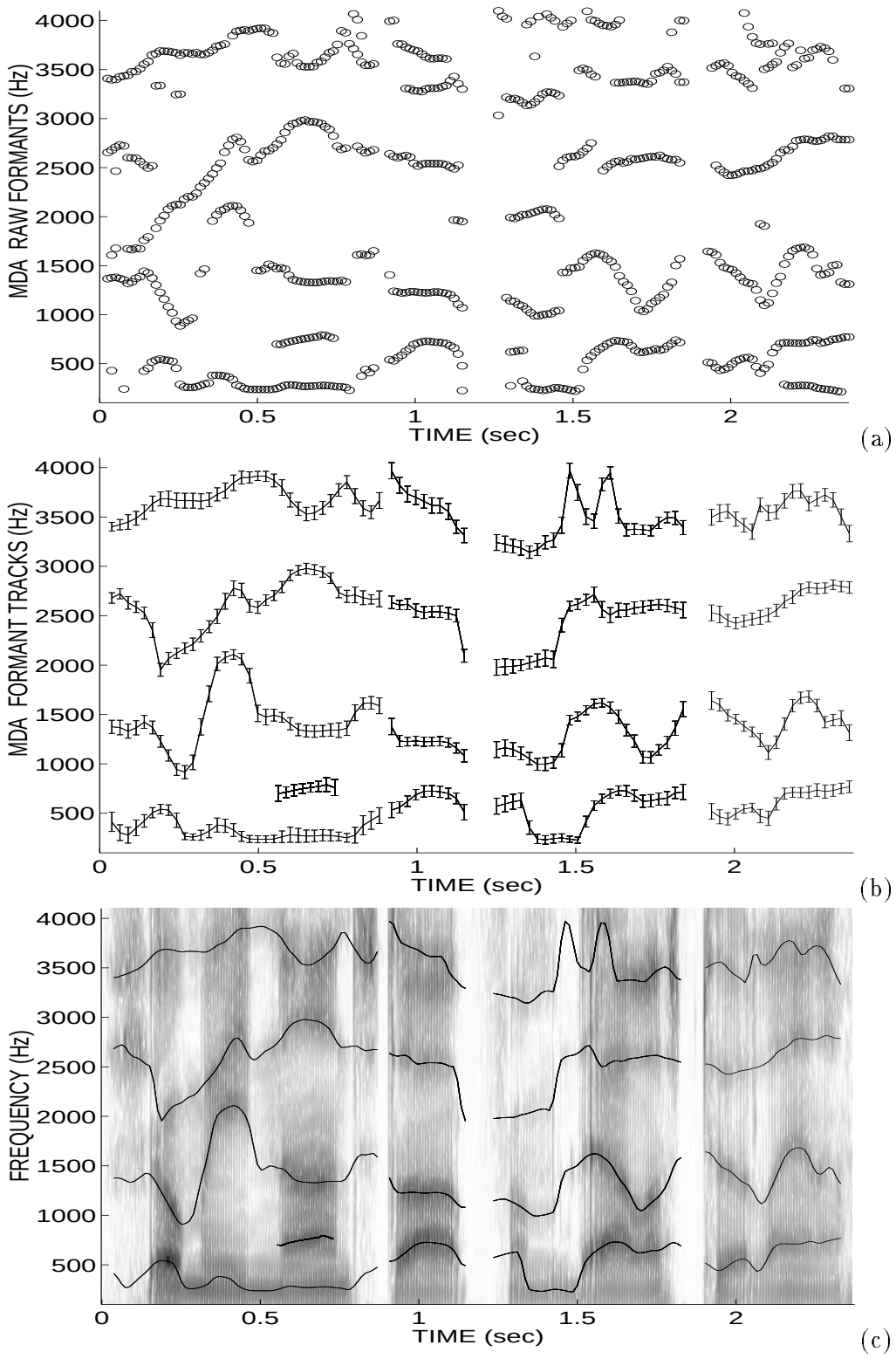

Figure 5: MDA formant tracking on the speech signal of Fig. 3(a): (a) Raw formant estimates, (b) Formant tracks: frequency and bandwidth (the bandwidth is equal to the length of the "error bar" centered at the formant frequency) and (c) Formant tracks superimposed on the speech spectrogram.

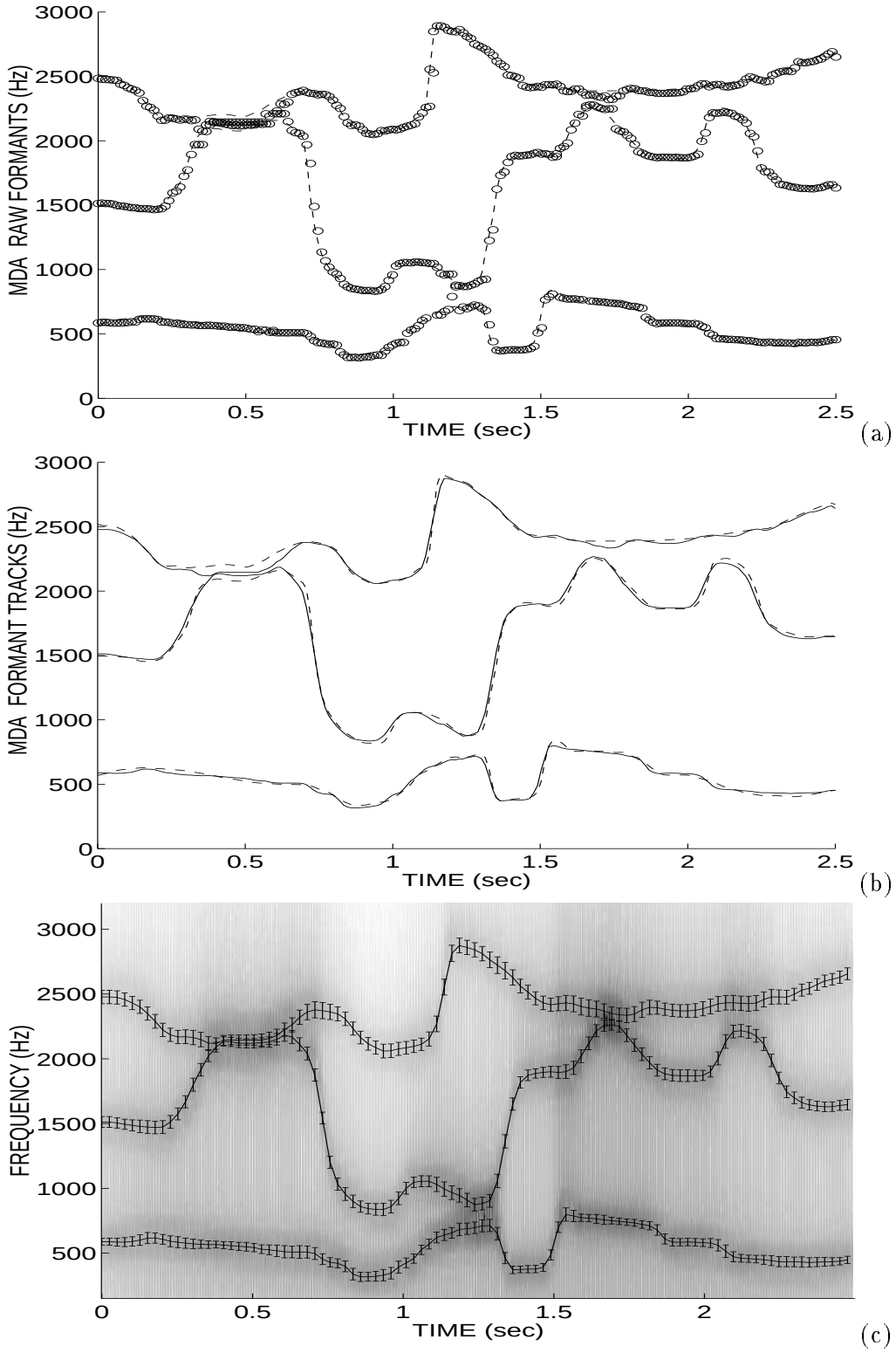

Figure 6: MDA formant tracking on synthetic speech: (a) Raw formant estimates. (b) Formant tracks: computed (solid line) vs. actual (dotted line). (c) Formant tracks superimposed on the speech spectrogram (formant bandwidths shown as "error bars" around the formant tracks).

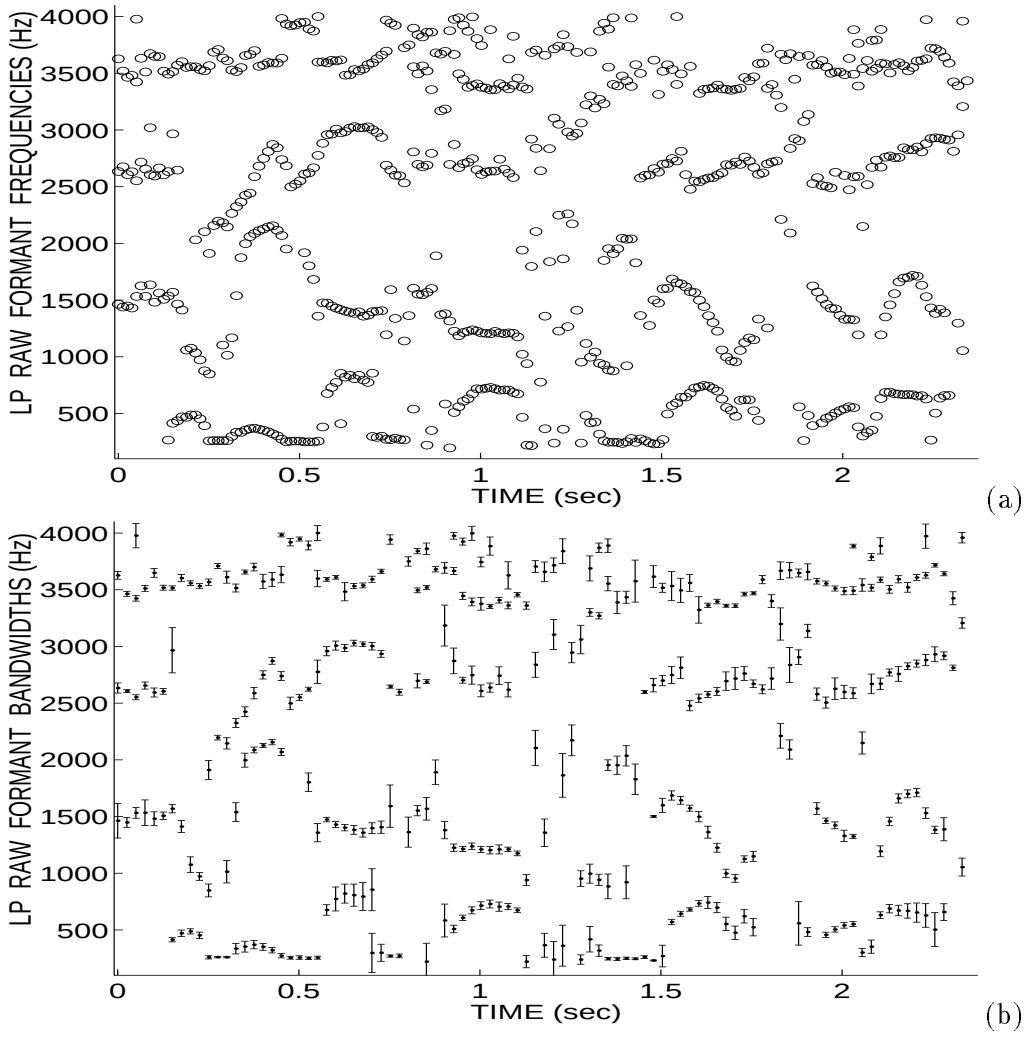

Figure 7: LP raw formant frequency (a) and bandwidth (shown as "error bars", scaled up 4 times) (b) estimates for the speech signal shown in Fig. 3(a); LP analysis order is 12, preemphasis is 0.5, window size is 25 msecs updated every 12.5 msecs.