Proc. of Intl. Conf. Acoustic, Speech and Signal Proc., Detroit, Michigan, May 1995.

SPEECH FORMANT FREQUENCY AND BANDWIDTH TRACKING USING MULTIBAND ENERGY DEMODULATION

Alexandros Potamianos and Petros Maragos

School of E.C.E., Georgia Institute of Technology, Atlanta, GA 30332-0250

ABSTRACT

In this paper, the AM–FM modulation model and a multiband analysis/demodulation scheme is applied to speech formant frequency and bandwidth tracking. Filtering is performed by a bank of Gabor bandpass filters. Each band is demodulated to amplitude envelope and instantaneous frequency signals using the energy separation algorithm. Short-time formant frequency and bandwidth estimates are obtained from the instantaneous amplitude and frequency signals and their merits are presented. The estimates are used to determine the formant locations and bandwidths. Performance and computational issues (frequency domain implementation) are discussed. Overall, the multiband demodulation approach to formant tracking is easy to implement, provides accurate formant frequency and realistic bandwidth estimates, and performs well in the presence of nasalization.

1. INTRODUCTION

Formant tracking is an old problem that has received much attention lately, mainly because of the deficiencies of the well established algorithms. Most formant tracking algorithms are based on linear prediction (LP) analysis 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. Alternatively, we propose a multiband demodulation approach to speech analysis in the framework of the AM–FM modulation model that overcomes these problems.

Motivated by several nonlinear and time-varying phenomena during speech production Maragos, Quatieri and Kaiser [4] proposed an AM–FM modulation model that represents a single speech resonance $R(t)$ as an AM–FM signal

$$R(t) = a(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \qquad (1)$$

where f_c 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 $f_i(t) = f_c + q(t)$. 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.

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

To obtain the amplitude envelope $|a(t)|$ and the instantaneous frequency $f_i(t)$ 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 spectrum before demodulation can be performed. These two steps of speech analysis will be referred to as *multiband demodulation* [1]. Multiband demodulation yields rich time-frequency information.

In section 3, we propose and compare two short-time estimates of the average frequency of a speech (frequency) band $R(t)$: the *mean instantaneous frequency*, which has been used for formant tracking in [3] and the *mean amplitude weighted instantaneous frequency*, a time-domain equivalent of the first central spectral moment [2]. Next, 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 accurate formant tracks and realistic formant bandwidth estimates, is easy to implement both in the time and frequency domain, and avoids most of the drawbacks of LP-based formant trackers.

2. DEMODULATION ALGORITHMS

A speech resonance $R(t)$ is extracted from the speech spectrum by a Gabor bandpass filter with impulse response $h(t) = \exp(-\alpha^2 t^2)\cos(2\pi f_c t)$, where f_c is the center frequency of the Gabor filter (chosen equal to the formant frequency) and α is the bandwidth parameter.

The *energy separation algorithm* (ESA) was developed in [4] to demodulate a speech resonance $R(t)$ into amplitude envelope $|a(t)|$ and instantaneous frequency $f_i(t)$ signals. The ESA is based on the Teager–Kaiser energy tracking operator $\Psi[s(t)] = [\dot{s}(t)]^2 - s(t)\ddot{s}(t)$, where $\dot{s} = ds/dt$. The ESA frequency and amplitude estimates are

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

Similar algorithms exist in discrete time [4, 5].

An alternative way to obtain $|a(t)|$ and $f_i(t)$ estimates is through the *Hilbert transform demodulation* (HTD), as the modulus and phase derivative of the Gabor analytic signal. The HTD can be implemented in the frequency domain (as an 90^o phase splitter). In general, the ESA has the advantages over HTD of being simple, computationally efficient, and of having excellent time resolution [7].

3. FORMANT FREQUENCY AND BANDWIDTH SHORT-TIME ESTIMATES

Simple short-time estimates F_1 and B_1 for the frequency and bandwidth of a formant candidate, respectively, are the mean and standard deviation of the instantaneous frequency signal $f_i(t)$, i.e.,

$$F_1 = \frac{1}{T} \int_{t_0}^{t_0+T} f_i(t)\, dt \tag{3}$$

$$[B_1]^2 = \frac{1}{T} \int_{t_0}^{t_0+T} (f_i(t) - F_1)^2\, dt \tag{4}$$

where t_0 and T are the start and duration of the analysis frame respectively. Alternative estimates can be found from the 1st and 2nd moments of $f_i(t)$ using the square amplitude $[a(t)]^2$ as weight

$$F_2 = \frac{\int_{t_0}^{t_0+T} f_i(t)\, [a(t)]^2\, dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{5}$$

$$[B_2]^2 = \frac{\int_{t_0}^{t_0+T} [(\dot{a}(t)/2\pi)^2 + (f_i(t) - F_2)^2 [a(t)]^2] dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{6}$$

where the additional term $(\dot{a}(t)/2\pi)^2$ in B_2 accounts for the amplitude modulation contribution to the bandwidth [2].

Consider the sum $x(t)$ of two sinusoids with constant frequencies f_1, f_2 and time-varying amplitudes $a_1(t)$, $a_2(t)$

$$x(t) = a_1(t) \cos[2\pi f_1 t] + a_2(t) \cos[2\pi f_2 t] \tag{7}$$

The amplitude envelope $|a(t)|$ and instantaneous frequency $f_i(t)$ signals (obtained from HTD) are

$$|a(t)| \approx (\, a_1^2 + a_2^2 + 2\, a_1\, a_2\, cos[\Delta\omega\, t]\,)^{\frac{1}{2}} \tag{8}$$

$$f_i(t) \approx (a_1^2 f_1 + a_2^2 f_2 + a_1 a_2 (f_1 + f_2)\, cos[\Delta\omega\, t])/a^2 \tag{9}$$

where $\Delta\omega = 2\pi(f_1 - f_2)$. The short time frequency estimates F_1 and F_2 are (depending on the frame boundaries)

$$F_1 \approx \begin{cases} f_1, & a_1 > a_2 \\ f_2, & a_1 < a_2 \end{cases} \qquad F_2 \approx \frac{a_1^2\, f_1 + a_2^2\, f_2}{a_1^2 + a_2^2} \tag{10}$$

i.e., F_1 locks on the harmonic with the larger amplitude, while F_2 provides a (square amplitude) weighted mean frequency. These results hold also when $|a(t)|$, $f_i(t)$ are computed via ESA (with slight modifications when $a_1 \approx a_2$).

In Fig. 1 (a),(b) we display the short-time frequency estimates F_1 and F_2 computed via the ESA (x) and the HTD (o) for the signal in (7) with $f_1 = 1.5$ kHz, $f_2 = 1.7$ kHz, $a_1(t) = 10\, t$, $a_2(t) = 1 - 10\, t$ and $t \in [0, 0.1]$ sec. The amplitudes of the two sinusoids are antagonizing so that for the first 50 msec the sinusoid f_2 is dominant while for the remaining 50 msec f_1 dominates. We can see that F_1 locks on the sinusoid with the greater amplitude (greater amplitude frequency product for the ESA), while F_2 being the mean (square amplitude) weighted frequency, provides a more 'intuitive' short-time frequency estimate.

For a sum of more than two (AM–FM) sinusoids: $F_2 \approx (\sum_n a_n^2\, f_n) \,/\, (\sum_n a_n^2)$, i.e., each frequency is weighted with the square amplitude. In general, the behavior of F_1 is complicated. If there are only one or two prominent sinusoids/harmonics in the spectrum though, F_1 will lock on the frequency of the sinusoid with the greatest amplitude.

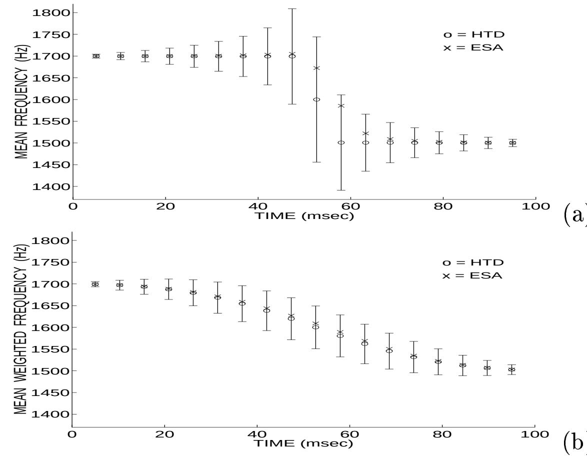

Figure 1: Short-time frequency and bandwidth (error bars) estimates for a sum of antagonizing AM sinusoids: (a) F_1, B_1, (b) F_2, B_2 (window 10 msec, updated every 5 msec).

A speech resonance can be thought of as the sum of a few slowly time-varying (AM–FM) sinusoids, one for each harmonic. Thus, F_1 has the tendency to lock on the harmonic with the greatest amplitude in the formant band, while F_2 weights the frequency of each harmonic with its square amplitude. In general, F_2 provides a more accurate formant frequency estimate; F_1 can be superior, however, when the bandpass filter used to extract the formant is not placed exactly on the spectral resonance (e.g. off by 100-200 Hz). Finally, there are cases where no major formant harmonics are 'inside' the bandpass filter; there the behavior F_1 is unpredictable (and thus unstable). Overall, the F_1 estimate is computationally simple, and converges faster to the formant frequency when the bandpass filter is 'misplaced'. The weighted frequency estimate F_2 is more robust for low energy bands (i.e., spectral valleys) and provides more accurate formant frequencies.

Similarly, B_2 bandwidth estimates are more robust than B_1 estimates. For example, in Fig. 1(a), (b) we display B_1 and B_2 (computed via HTD) for the sum of two sinusoids in Eq. (7). 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_1 takes unnaturally large values.

In [2], the (square amplitude) weighted estimates F_2 and B_2 are shown to be time domain equivalents of the first and second central spectral moments of the signal; this offers an alternative way of computing F_2 and B_2 in the frequency domain (see also section 5).

Overall, the HTD and the ESA provide similar estimates F, B, because of the smoothing effects of bandpass filtering. The ESA, however, has lower computational complexity [7]. When the center frequency of the bandpass filter approaches the pitch frequency, though, the HTD produces smoother B_2 estimates than the ESA. Thus, if formant bandwidth estimates are needed in the 0-500 Hz frequency range (first formant), the HTD should be preferred; otherwise, for higher formants, the ESA should be used for computational efficiency.

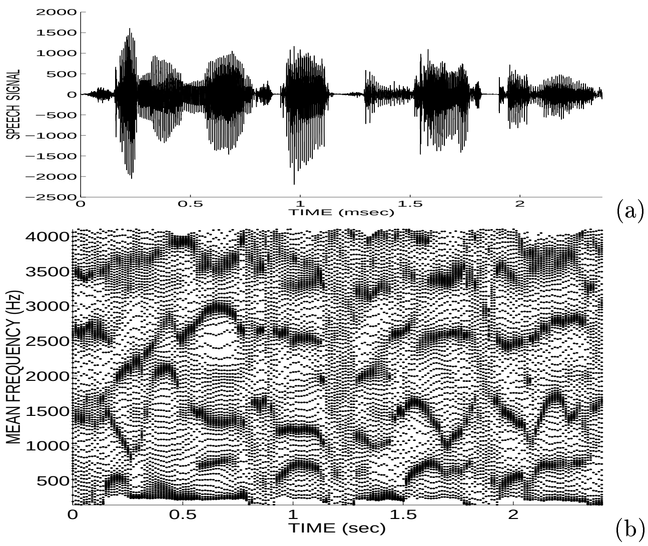

Figure 2: (a) Speech signal: 'Show me non-stop from Dallas to Atlanta' and (b) short-time frequency estimate $F_2(t, f)$ for the output of 80 Gabor filters (f spanning 200 to 4200 Hz) v. time (pyknogram).

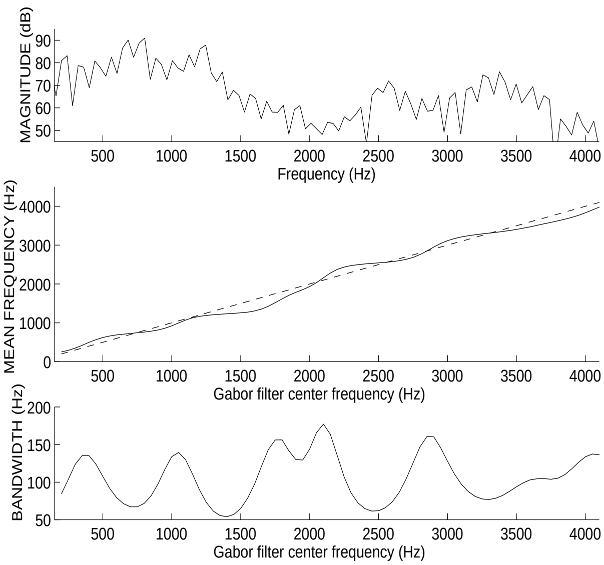

Figure 3: The short-time Fourier transform, the frequency $F_2(f)$ and bandwidth $B_2(f)$ estimates vs. the center frequencies f of the Gabor filters (25 msec frame).

4. MULTIBAND DEMODULATION FORMANT TRACKING ALGORITHM

Next, we introduce the parallel multiband filtering and demodulation algorithm for formant tracking. The speech signal is filtered through a bank of Gabor bandpass filters, uniformly spaced in frequency. The amplitude envelope $|a(t)|$ and instantaneous frequency $f_i(t)$ signals are estimated for each Gabor filter output. Short-time frequency $F_2(t, f)$ and bandwidth $B_2(t, f)$ estimates are obtained from the instantaneous amplitude and frequency signals (Eqs. (5), (6)), for each speech frame located around time t and for each Gabor filter of center frequency f. The time-frequency distributions thus obtained have time resolution equal to the step of the short-time window (typically 10 msec) and frequency resolution equal to the center frequency difference of two adjacent filters (typically 50 Hz).

In Fig. 2(b), we plot the value of the short-time frequency estimates $F_2(t, f)$ for each and every frequency band (centered at frequency f) vs. time t for the sentence in (a), i.e., unlike a typical time-frequency plot, here, the y-axis represents the range rather than the domain of F_2. Note the dense concentration of points (frequency estimates) around the formants tracks. The plot density plays the role that the Fourier magnitude plays in a speech spectrogram, so we refer to this time-frequency representation as the *speech pyknogram*. The pyknogram displays both the formant positions (and bandwidths) and the location of the spectral zeros (low density areas).

In Fig. 3, we show the frequency $F_2(f)$ and bandwidth $B_2(f)$ estimates for a single analysis frame, vs. the center frequency of the Gabor filters f. We have observed that bandwidth B_2 minima consistently indicate the presence of formants.

In order to determine robust raw formant estimates for a frame of speech we search for points where $F_2(f)$ and the Gabor filter center frequency f are equal (i.e., $F_2(f) = f$, or in Fig. 3 the points where the solid line meets the dotted one) and $dF_2(f)/df < 1$. In addition, there are cases where a weak formant is 'shadowed' by a strong neighboring one; then $F_2(f)$ approaches the line f without reaching it. Thus, we also search for points where $F_2(f) - f$ has local maxima and $F_2(f) < f$. These points are also considered formant estimates if the difference $f - F_2(f)$ is less than a threshold (typically 50 Hz). Finally, we improve the accuracy of the formant estimates by linear interpolation.

An alternative way of determining the raw formants is by determining the dense regions of Fig. 2(b), i.e., determining the minima of $dF_2(f)/df$. This approach may produce better formant estimates when two formant tracks are close together. However, a threshold must be provided for the plot density, which makes this approach less reliable.

In Fig. 4(a), we display the raw formant estimates for the sentence of Fig. 2(a). A 3-point binomial smoother is applied on $F_2(t, f)$ in the time domain before the raw formant estimates are computed. In Fig. 4(b) the formant tracks (frequency and bandwidth) are shown. The decision algorithm used is similar to linear prediction (LP) based formant tracking algorithms [6], with special care taken for nasals sounds (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 formants tracks are well separated in frequency and well defined). Next, we determine if a 'nasal formant' is present between the first and the second formant. Finally, the formant tracks between anchor segments are filled. Formant bandwidths are obtained from the B_2 estimate.

Most formant tracking algorithms are based on a short-time linear prediction analysis. LP-based formant trackers encounter problems with nasals and nasalized vowels. In

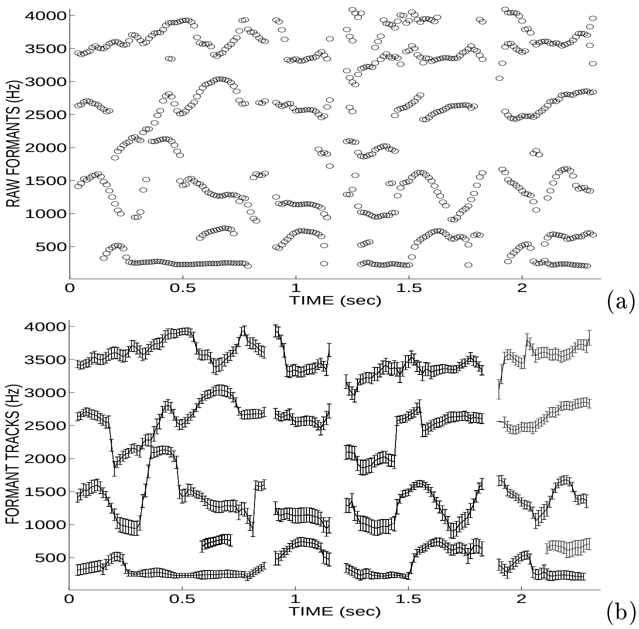

Figure 4: (a) Raw formant estimates and (b) Formant tracks: frequency and bandwidth (error bars).

addition, the formant accuracy is affected by the preemphasis and the harmonic structure of the spectrum (LP tends to 'ride with the harmonics' as does the F_1 frequency estimate). Finally, LP does not provide accurate formant bandwidth estimates. By using the multiband demodulation algorithm one avoids these problems. Overall, the multiband demodulation 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 (it tracks an extra 'nasal formant') and provides realistic formant bandwidth estimates.

We saw in section 3 that the choice of F_1 vs. F_2 is the choice between fast convergence and robust raw formant estimates. Since convergence is not the issue here, we prefer to use the more reliable weighted estimate F_2. Note, though, that when the frequency axis is poorly sampled (i.e., when only a few Gabor filters are used), F_1 can produce better results than F_2, since F_1 provides good formant estimates even when the Gabor filter is not centered exactly on the formant frequency.

5. DISCUSSION

In [3] an iterative algorithm (iterative ESA) is proposed for formant tracking: initial formant estimates are refined through an iterative filtering/demodulation scheme. For the iterative ESA, the F_1 frequency estimate is preferred to F_2, since F_1 increases substantially the convergence speed to a formant. In general, the iterative ESA and the parallel multiband algorithm (proposed above) yield similar results. In the case where two formants tracks merge, though, the multiband algorithm performs better due to its more flexible decision about raw formant values.

In the multiband formant tracking algorithm we use a bank of uniformly spaced (in frequency) Gabor filters. Alternatively, a Gabor wavelet bank with constant logarithmic spacing can be used, which gives uniform performance (for ESA demodulation) across channels [1]. Logarithmic spacing (over 1 kHz) is also compatible with the formant frequency perceptual resolution (limens) of the ear. In [3], using logarithmic spacing has improved the performance of the iterative ESA formant tracker.

We mentioned in section 3 that the estimates F_2 and B_2 can be computed in the frequency domain as the first and second spectral moments (computation via 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_2 and B_2 estimates computed in the frequency domain take similar values to their time domain equivalents when adequately 'long' FFT implementation is used (512 or 1024 point FFT gives good results).

Finally, one may use multiband demodulation for spectral zero tracking. In Fig. 2(b), zeros manifest themselves as areas of low plot density (e.g., for nasalized sounds an antiformant can be observed between the second and the third formant track). More work is on the way for anti-formant tracking using the multiband ESA.

Overall, the multiband demodulation formant tracker produced very promising results, which suggest that the modulation model and the demodulation algorithms are a useful alternative modeling/analysis approach to speech processing.

6. REFERENCES

[1] A. C. Bovic, P. Maragos, and T. F. Quatieri, "AM–FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3245–3265, Dec. 1993.

[2] L. Cohen and C. Lee, "Instantaneous bandwidth," in *Time Frequency Signal Analysis - Methods and Applications* (B. Boashash, ed.), London, 1992.

[3] H. M. Hanson, P. Maragos, and A. Potamianos, "A system for finding speech formants and modulations via energy separation," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 436–443, July 1994.

[4] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3024–3051, Oct. 1993.

[5] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. Signal Proc.*, vol. 41, pp. 1532–1550, Apr. 1993.

[6] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. ASSP–22*, pp. 135–141, Apr. 1974.

[7] 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.