Proc. of VII European Signal Processing Conference, Edinburgh, Scotland, Sept. 1994.

APPLICATIONS OF SPEECH PROCESSING USING AN AM–FM MODULATION MODEL AND ENERGY OPERATORS

Alexandros Potamianos and Petros Maragos

School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA 30332-0250, U.S.A.

ABSTRACT

A recent speech modulation model represents each resonance (formant) as an AM–FM signal. Resonances are demodulated into instantaneous amplitude and frequency signals using the energy separation algorithm. We present three applications of these ideas (1) a multiband parallel demodulation formant tracking algorithm, (2) an AM-FM vocoder which codes the amplitude and frequency components of each formant band, and (3) the energy spectrum which yields a non-parametric smooth spectral envelope.

1. INTRODUCTION

Recently, the importance of modulations in speech resonances has come to the attention of the speech community. Motivated by several nonlinear and time-varying phenomena during speech production Maragos, Quatieri and Kaiser [5] 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 $f_i(t) = f_c + q(t)$. 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.

The *energy separation algorithm* (ESA) was developed in [5] to demodulate a speech resonance $R(t)$ into amplitude envelope $|a(t)|$ and instantaneous frequency $f_i(t)$ signals. The ESA is based on an energy-tracking operator introduced by Teager and Kaiser [4], which tracks the energy of the source producing an oscillation signal $s(t)$ and is defined as

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

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

$$f_i(t) \approx \frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}}, \qquad |a(t)| \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \qquad (3)$$

Similar equations and algorithms exist in discrete time [5, 6]. The ESA is simple, computationally efficient, and has excellent time resolution [7].

The AM–FM modulation model, the energy operator and the ESA have proven to be useful tools in several speech analysis and synthesis applications. The applications presented in this paper are (1) a parallel *formant tracking* algorithm using the multi-band ESA [2], (2) an *AM–FM modulation vocoder*, which extracts the formant bands from the spectrum, demodulates them and codes the instantaneous amplitude and frequency signals, and (3) the *energy spectrum*, a smooth spectral envelope of the speech signal.

2. FORMANT TRACKING

In [3] an iterative ESA scheme is used for formant tracking. Here, we propose a *multi-band parallel demodulation algorithm*. The speech signal is filtered through a bank of Gabor band-pass filters with fixed center frequencies and bandwidths. The Gabor filters are uniformly spaced in frequency and have constant bandwidth. Next, the amplitude envelope $|a(t)|$ and instantaneous frequency $f_i(t)$ are estimated for each filtered signal. Short-time frequency $F(t, f)$ and bandwidth $B(t, f)$ estimates are obtained from the instantaneous amplitude and frequency signals, 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 (shift) of the short-time window (typically 10 msec) and frequency resolution equal to the center frequency difference of two adjacent filters (typically 50 Hz). F and B are the features used for raw formant estimation and formant tracking.

To demodulate the filtered signals into their amplitude envelope $|a(t)|$ and instantaneous frequency $f(t)$ components one may use two alternative algorithms: the energy separation algorithm (ESA) or the Hilbert transform demodulation (HTD). The ESA is simpler,

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

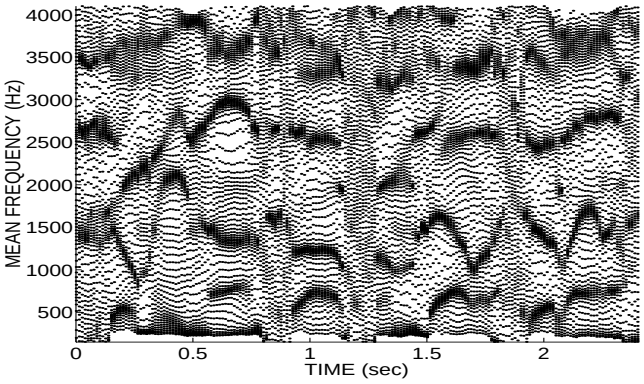

Figure 1: Short-time frequency estimate $F_2(t,f)$ for the output of 80 Gabor filters (center frequency f spanning 200 to 4200 Hz) vs. time, for the sentence 'Show me non-stop from Dallas to Atlanta'.

computationally more efficient and has better time resolution, but its performance deteriorates as the center frequency of the Gabor filter approaches the pitch frequency. In that case we have found that the HTD (implemented via FFT) produces smoother estimates, but at a higher computational complexity. The two approaches are compared in [7].

Simple short-time estimates F_1 and B_1 for the frequency F and bandwidth B of a formant candidate, respectively, are the frequency and the standard deviation of the instantaneous frequency signal, i.e.,

$$F_1(t_0,f) = \tfrac{1}{T}\ \int_{t_0}^{t_0+T}\ f_i(t)\ dt$$

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

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

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

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

The estimates F_1, B_1 are conceptually simple and easy to compute, while F_2, B_2 (which we use henceforth) are more robust (this property is important in a iterative scheme [3, 7]). If only frequency estimates $F(t,f)$ are needed, the ESA is used for computationally efficient demodulation. Smoother bandwidth estimates $B(t,f)$ for frequencies f below 1 kHz have been obtained via the HTD.

In Fig. 1, we plot the short-time frequency estimate $F_2(t,f)$ for all bands vs. time. Note the dense concentration of estimates around the frequency tracks. The plot

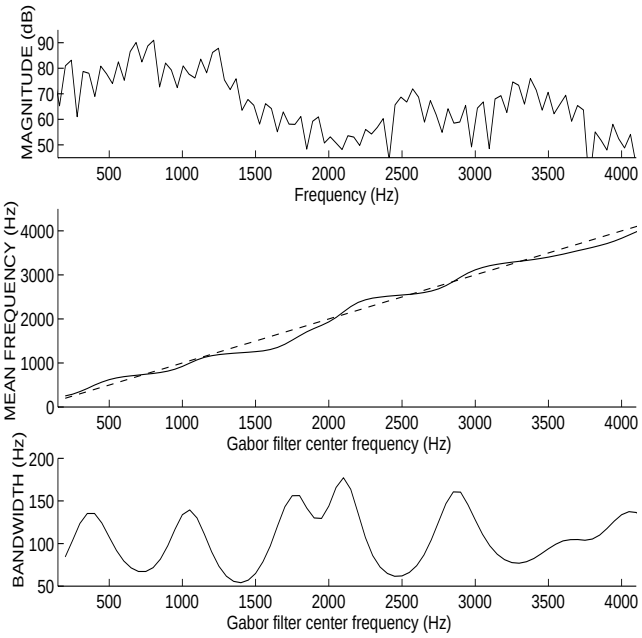

Figure 2: The short-time Fourier transform, the frequency $F_2(f)$ and bandwidth $B_2(f)$ estimates vs. the center frequencies f of the Gabor filters, for a 25 msec frame of speech.

density plays the role that the Fourier magnitude plays in a speech spectrogram. In Fig. 2, we show frequency $F_2(f)$ and bandwidth $B_2(f)$ estimates for a single analysis frame. 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. 2 the points where the solid line meets the dotted one) and $dF_2(f)/df < 0$. 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.

In Fig. 3(a), we display the raw formant estimates for the sentence of Fig. 1. 3-point binomial smoothing is performed on $F_2(t,f)$ in the time domain before the raw formant estimates are computed. In Fig. 3(b) the formant tracks (frequency and bandwidth) are shown. The decision algorithm used is similar to LPC-based formant tracking algorithms, with special care taken for nasals sounds (a 'nasal formant' between F1-F2 is allowed to be born and to die). Formant bandwidths are obtained from B_2.

The multi-band parallel demodulation formant tracking algorithm has the attractive features of being con-