SPEECH ANALYSIS AND SYNTHESIS USING AN AM–FM MODULATION MODEL

Alexandros Potamianos † *and Petros Maragos* *

† AT&T Labs–Research, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, U.S.A.
* Institute for Language & Speech Processing, Margari 22, Athens 11525, Greece
and School of E.C.E, Georgia Institute of Technology, Atlanta, GA 30332, USA.

ABSTRACT

In this paper, the AM–FM modulation model is applied to speech analysis, synthesis and coding. The multiband demodulation pitch tracking algorithm is proposed that produces smooth and accurate fundamental frequency contours. The AM–FM modulation vocoder represents speech as the sum of resonance signals modeled by their amplitude envelope and instantaneous frequency signals. Efficient modeling and coding (at 4.8-9.6 kbits/sec) algorithms are proposed for the amplitude envelope and instantaneous frequency signals. Amplitude and frequency modulations of the speech resonances are shown to be perceptually important for natural speech synthesis.

1. INTRODUCTION

Despite the well-known existence of nonlinear and time-varying phenomena during speech production the linear source-filter model is extensively used as the foundation of speech modeling. Deviations from these linear assumptions are mathematically modeled, often with little concern about the underlying physical phenomena. Such models can accurately reproduce and synthesize speech for some speakers using concatenative methods, but they are not equally successful in transforming speaker characteristics and speaking styles in a controlled way.

Motivated by nonlinear and time-varying phenomena during speech production and the need for a better understanding of the speech production process Maragos, Kaiser and Quatieri [3] proposed a nonlinear model that describes a speech resonance as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure

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

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 $f(t) = f_c + q(t)$. The speech signal $s(t)$ is modeled as the sum $s(t) = \sum_{k=1}^{K} r_k(t)$ of K such AM–FM signals, one for each formant.

In this paper, the AM–FM modulation model is applied to speech analysis, synthesis and coding. Further, an attempt is made to understand the perceptual importance of amplitude and frequency modulations in speech resonances. The organization of this paper is as follows: First multiband demodulation is introduced, the analysis tool used extensively in this paper. In Section 2.3 an application of the AM–FM modulation model and multiband demodulation analysis to the problem of fundamental frequency estimation is presented. In Section 3, the AM–FM analysis/synthesis system is presented and efficient coding algorithms are proposed for the amplitude and frequency modulating signals of each resonance. Finally, the perceptual importance of modulations is discussed in Section 4.

This work was performed while the authors were with the School of E.C.E, Georgia Institute of Technology, Atlanta, GA 30332, USA. It was partially supported by the US National Science Foundation under Grants MIP-9396301 and MIP-9421677.

2. SPEECH ANALYSIS

2.1. Multiband Demodulation Analysis

A speech resonance (or, in general, speech frequency band) signal $r(t)$ is extracted from the speech signal $x(t)$ through bandpass filtering. A real Gabor filter is used for this purpose. The amplitude envelope $|a(t)|$ and the instantaneous frequency $f(t)$ signals are obtained by applying the energy separation algorithm (which is an AM–FM demodulation algorithm) on the speech resonance signal $r(t)$. A formal discussion on using Gabor wavelets for multiband demodulation analysis (MDA) can be found in [2].

The *energy separation algorithm* (ESA) is based on the nonlinear differential Teager–Kaiser energy operator [3]. The energy operator tracks the energy of the source producing an oscillation signal $r(t)$ and is defined as $\Psi[r(t)] = [\dot{r}(t)]^2 - r(t)\ddot{r}(t)$ where $\dot{r} = dr/dt$. The ESA frequency and amplitude estimates are

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

Similar equations and algorithms exist in discrete time. An alternative way to estimate $|a(t)|$, $f(t)$ is the Hilbert transform demodulation (HTD) algorithm, i.e., as the modulus and the phase derivative of the Gabor analytic signal (see [5] for ESA, HTD comparison).

2.2. Formant Tracking

In [6], multiband demodulation analysis was applied to formant tracking. Formant frequency estimates (F_u, F_w) were proposed

$$F_u = < f(t) >_T\ , \qquad F_w = < a^2(t)\ f(t) >_T \qquad (3)$$

where $|a(t)|$, $f(t)$ are the amplitude envelope and the instantaneous frequency signals of resonance signal $r(t)$, and $< . >_T$ denotes averaging over a window of duration T. It was shown that F_w is approximately equal to the weighted

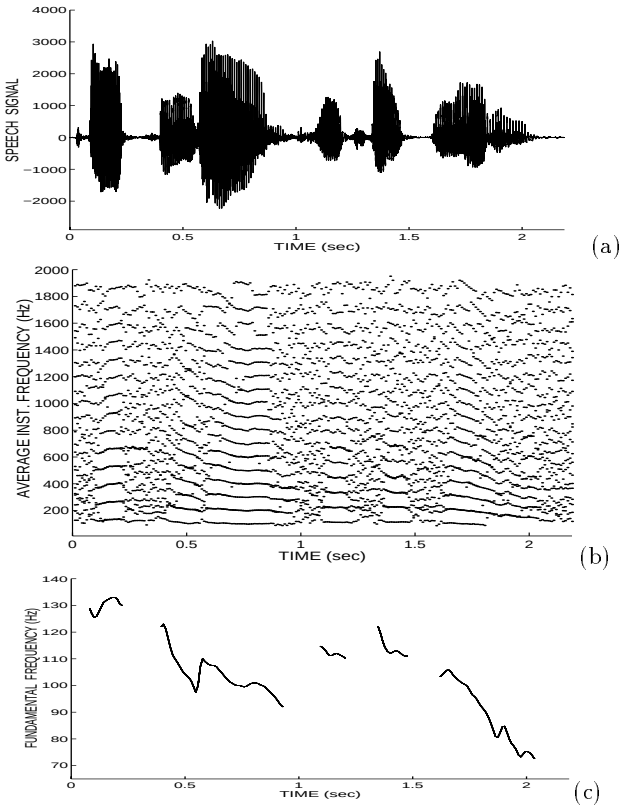

Figure 1: (a) Speech signal: "Cats and dogs each hate the other." (b) Time-Frequency average instantaneous frequency distribution. (c) MDA fundamental frequency estimate.

average of the harmonic frequencies, with weights the harmonic amplitudes squared, while F_u (under certain constraints) equals the frequency of the most prominent harmonic in the frequency band. Thus, F_w was selected as the formant frequency estimate.

The MDA formant tracking algorithm is outlined next. The speech signal is filtered through a *fixed* bank of Gabor bandpass filters, uniformly spaced in frequency (typical effective RMS Gabor filter bandwidth is 400 Hz and spacing is 50 Hz). Next, the amplitude envelope $|a(t)|$ and instantaneous frequency $f(t)$ signals are estimated for each Gabor filter output using the ESA. The short-time weighted instantaneous frequency $F_w(t,\nu)$ is computed for each speech frame located around time t (every 10 msec) and for each Gabor filter centered at frequency ν. The time-frequency distribution $F_w(t,\nu)$ is used to determine the raw formant tracks. Finally, the tracks are refined using global continuity constraints.

2.3. Fundamental Frequency Estimation

As discussed in [6], the short-time average of the instantaneous frequency F_u is an accurate estimate of the most dominant frequency in the signal's spectrum (for narrowband signals).[1] Next, a multiband demodulation pitch tracking algorithm is proposed using F_u as a harmonic frequency estimate.

Similarly to MDA formant tracking, the speech signal is filtered through a bank of Gabor bandpass filters (typical effective RMS Gabor filter bandwidth is 200 Hz and the approximate spacing is 100 Hz following a mel frequency scale) and then each bandpass signal is demodulated to amplitude envelope and instantaneous frequency signals. The short-time average of the instantaneous frequency signal F_u is computed and is used as an estimate of the most prominent harmonic in each band. The time-frequency distribution of $F_u(t,\nu)$ is shown in Fig. 1(b) for a sentence from the TIMIT database. The harmonic tracks are clearly visible. The fundamental frequency of a voiced speech segment is determined from the minimization of the error sum $E(\mathrm{F0})$ over all possible fundamental frequency candidates F0

$$E(\mathrm{F0}) = \frac{1}{\mathrm{F0}} \sum_{n=1}^{N} \mid F_u(\nu_n) - \lfloor \frac{F_u(\nu_n)}{\mathrm{F0}} + 0.5 \rfloor \, \mathrm{F0} \mid \qquad (4)$$

where ν_n is the center frequency of the nth Gabor filter in the filterbank, N is the total number of filters and $F_u(\nu_n)$ is the average instantaneous frequency for the band centered at frequency ν_n. Optionally, weighting factors $\alpha(\nu_n)$ measuring the relative prominence of the estimated harmonic $F_u(\nu_n)$ can be added in the error functional E. In the error sum of Eq. (4), deviations of the phase slope estimate from the nearest multiple of the fundamental frequency candidate are penalized. The estimated fundamental frequency F0 provides the best match between the short–time harmonic estimates $\{n : F_u(\nu_n)\}$ and the fundamental frequency multiples $\{k : k\mathrm{F0}\}$ (sinusoidal pitch trackers use a similar "harmonic matching" approach [4]). The algorithm produces very detailed and smooth fundamental frequency contours as shown in the example of Fig. 1(c) for the speech signal in Fig. 1(a).

The pitch contours are filtered to correct few occurrences of "pitch-halving." Alternatively, a global error functional can be defined for each voiced region that explicitly penalizes pitch discontinuities. The global error E_G to be minimized over all possible pitch paths $\mathrm{F0}(t)$ is defined as

$$E_G = \int_{t_1}^{t_2} E[\mathrm{F0}(t)] \, dt + \lambda \int_{t_1}^{t_2} \left(\frac{d\mathrm{F0}(t)}{dt} \right)^2 dt \qquad (5)$$

for each voiced region $[t_1, t_2]$. E is the error criterion of Eq. (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 pitch estimates can be further refined (error < 1 Hz) with minimal increase in computational complexity by pitch-synchronous averaging of the instantaneous frequency signal $f(t)$ in a second pass of the pitch tracking algorithm. Specifically, we have shown that if the analysis window of duration T is a pitch period multiple the accuracy of the F_u estimate is

$$F_u = f_M + O(\epsilon^4) \,, \quad \epsilon = \max_{k \neq M}(a_k / a_M) \qquad (6)$$

where f_M is the most prominent harmonic in the spectrum band ($a_M = \max_k(a_k)$) and a_k is the amplitude of the kth harmonic f_k. Note that the error is at worst $O(\epsilon)$ for a fixed window duration T.

3. THE AM–FM MODULATION VOCODER

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 $s(t)$ along the formant tracks. The formant tracks are obtained from the

[1] Alternatively, the slope of the phase signal computed from linear regression can provide more noise-robust estimates (the phase signal is the cumulative sum of the instantaneous frequency signal)[5].

multiband demodulation formant tracking algorithm (see Section 2.2). Filtering is performed by a bank of Gabor filters with time-varying center frequencies that follow the formant tracks. Next, the resonance signals are demodulated to amplitude envelope $|a_k(t)|$ and instantaneous frequency $f_k(t)$ signals using the ESA. The information signals $|a_k(t)|$, $f_k(t)$ have typical bandwidths of 400–600 Hz and are decimated by a factor of 20:1 (for 16 kHz sampling frequency). Finally, the decimated information signals are modeled and coded (see next section). 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. The synthetic speech signal $\hat{s}(t)$ is the sum of the reconstructed formant bands. The block diagram of the AM–FM modulation analysis–synthesis system is shown in Fig. 2.

The AM–FM vocoder is related to the parallel formant vocoder, since both vocoders model the speech signal as a superposition of formant resonance 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. Further, by retaining the excitation–vocal tract coupling the AM–FM modulation model allows us more freedom to investigate nonlinear speech production phenomena not modeled by the source-linear filter model. Next, the amplitude envelope and instantaneous frequency signals are modeled and the perceptual importance of amplitude and frequency modulations in speech is investigated.

3.1. Modeling of the Modulation Signals

The amplitude envelope signals of different formants are highly correlated for voiced speech and have a specific structure. To exploit this structure a multipulse model [1] is used for modeling the amplitude envelope. The multipulse excitation signals for amplitude envelopes of different formant bands are expected to be coupled for voiced speech and loosely coupled for unvoiced speech.

The model used for the amplitude envelope is

$$a(n) = u(n) * g(n) * h_l(n), \quad u(n) = \sum_{k=1}^{K} b_k \delta(n - n_k) \quad (7)$$

where $u(n)$ is the excitation signal, $g(n)$ is the impulse response of a critically damped second order system and $h_l(n)$ is the demodulated impulse response of the filter used for extracting the corresponding resonance signal $r(t)$ (for a real Gabor filter $h_l(t) = exp(-\alpha t^2)$). The main reason for using a critically damped second order filter $g(n)$ is the inability of the unconstrained linear predictor to model the perceptually important information of the envelope signal $a(t)$. The impulse response of this *monoparametric* critically damped system $g(n)$ was found to be a good approximation to the amplitude envelope of real speech resonances for both the attack and the (exponential) decay portions of the signal. Finally, $h_l(n)$ was introduced in the amplitude envelope model of Eq. (7) to account for the distortion introduced in $a(t)$ from the (Gabor) bandpass filtering procedure. The pulse positions n_k are computed from the analysis-by-synthesis loop while the amplitudes b_k have a closed form solution [1, 5].

In Fig. 3(b) the amplitude envelope and the corresponding excitation signals (computed as described above) are shown for the first and second resonances of the speech signal in Fig. 3(a). Two to three pulses per pitch period are used to model the amplitude envelope signal. The excitation pulses at the beginning of each pitch period correspond to the primary excitation instants, while the rest model secondary excitations and nonlinear phenomena. Note that the primary pulse positions for F1 and F2 are very close.

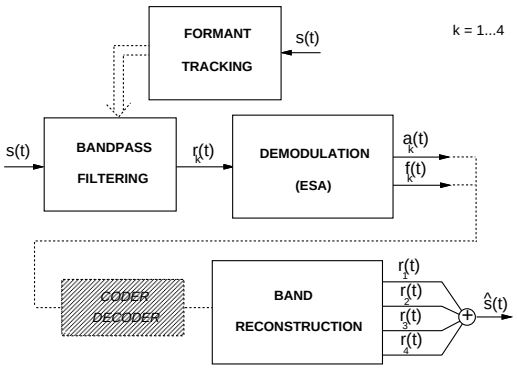

Figure 2: Block diagram of the modulation vocoder

The instantaneous frequency signal is modeled as the superposition of a slow- and a fast-varying component. The slow-varying component models the average formant frequency values and the fast-varying component models frequency variations around the formant frequency. A simple piece-wise linear model is assumed for the fast-varying frequency modulation component. Specifically, the instantaneous frequency takes different values for the open and closed phase of voicing. In Fig. 3(c), the (actual) instantaneous frequency signals and formant tracks (dashed) are shown for F1 and F2.

3.2. Coding of the Modulation Signals

To code the excitation signal $u(n)$ the pulses are classified into primary and secondary groups and each group is coded separately. The distances between consecutive primary pulses form a slowly time-varying contour ("pitch" contour) which can be efficiently coded. Similarly, the amplitudes of the excitation pulses form a "smooth" contour and are coded using PCM. Secondary pulse positions are coded relatively to the primary ones. Typically 3.5 to 6 kbits/sec are used to code the envelope signals.

The slow- and fast-varying components of the instantaneous frequency signals are coded separately. Formant tracks are decimated to 40 Hz and coded using PCM (abrupt formant transitions are also modeled). The phase at primary excitation instants and the frequency modulation index are sampled coarsely and coded for the first frequency band only (where FM is perceptually most important). Typically 1.3 to 3 kbits/sec are alloted to the instantaneous frequency signals.

The AM–FM vocoder produces very natural speech at 4.8 to 9.6 kbits/sec. Yet, the modeling of the very low (0-200 Hz) and very high ($>$ 5 kHz) frequency regions is inadequate. Additional work is needed to improve the proposed coding and modeling algorithms.

4. DISCUSSION

In this section, the perceptual importance of amplitude and frequency modulations is discussed. Further, alternative ways of modeling the amplitude and frequency modulation patterns are proposed.

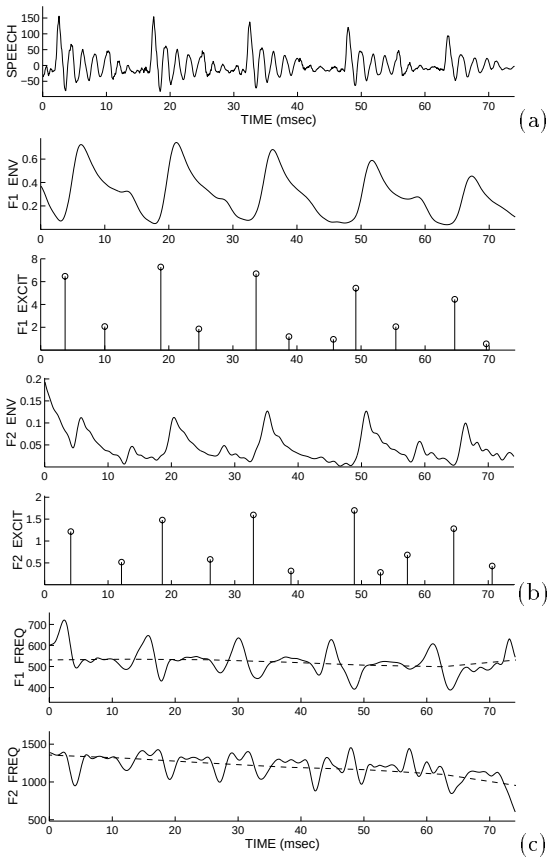

Figure 3: (a) Speech signal, phoneme /ow/ from "zero". (b) Amplitude envelope and multipulse excitation signals for the first and second resonances. (c) Instantaneous frequency signals and formant tracks for F1, F2.

From informal listening tests it was verified that the amplitude and frequency modulation of speech resonances are perceptually important for producing natural sounding speech. From preliminary experiments on synthetic speech and sentences from the TIMIT database using the AM–FM modulation vocoder it was determined that amplitude modulations convey both phonemic and speaker-dependent information. For bandpassed synthetic speech (with only a single formant on average in the passband) it was shown that adding amplitude modulations can alter the perceived phonemic quality of the sound. The existence of complementary information in resonance modulations may be the main reason for the increased intelligibility of noise-corrupted natural speech vs. (identically corrupted) speech produced by a formant synthesizer.

The AM–FM analysis-synthesis system is a valuable tool for measuring modulations in speech resonances. Alternatively, one can investigate modulations in speech using a frequency domain model. Amplitude modulations appear in the DFT spectrum as a departure from the shape of the linear formant peak, e.g., as an asymmetric formant peak or a peak were certain harmonics have reduced amplitudes. A simple short-time model that can quantify such phenomena is the sinusoidal model [4] applied to the formant resonance signal, i.e., express the speech resonance signal as a superposition of sinusoids and quantify the modulation amount by the difference between the amplitudes of the sinusoids for an actual and synthetic speech resonance. Similar ideas have been discussed in [7]. The model can be further enhanced to account for time-varying modulation amounts. Frequency modulation is not clearly visible from the DFT of the signal. A sinusoidal model with modulated (time-varying) amplitudes in the analysis window could capture some of the frequency modulation phenomena. By combining sinusoidal and resonance modeling additional intuition can be gained in the physical significance of modulations in speech.

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. 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. 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 formants. More work is underway to quantify how such modifications affect the speech quality. An application area of the vocoder is text-to-speech (TTS) synthesis and speaker transformation.

5. CONCLUSIONS

The AM–FM modulation model and multiband demodulation were successfully applied to speech analysis. The multiband demodulation pitch tracking algorithm was proposed that produces smooth and accurate fundamental frequency contours. Efficient modeling and coding algorithms were proposed for the amplitude envelope and instantaneous frequency resonance signals of the AM–FM modulation vocoder. The vocoder produces natural speech at 4.8-9.6 kbits/sec. Amplitude and frequency modulations were shown to convey both phonemic and speaker-dependent information and to be perceptually important for producing natural sounding speech.

6. REFERENCES

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural–sounding speech at low bit rates," in *Proc. ICASSP'82*, pp. 614–617.

[2] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM–FM energy detection and separation in noise using multiband energy operators," *Trans. SP-41*, no. 12, pp. 3245–3265, Dec. 1993.

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

[4] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", in *Trans. on Signal Processing*, vol. 34, pp. 744–754, 1986.

[5] A. Potamianos, "Speech processing applications using an AM–FM modulation model", Ph.D. Thesis, Harvard University, Aug. 1995.

[6] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *Journal of the Acoustical Society of America*, vol. 99, pp. 3795–3806, June 1996.

[7] T. F. Quatieri, personal communication, 1997.