Speech Analysis and Synthesis Using an AM–FM Modulation Model

Alexandros Potamianos † and Petros Maragos *

† Bell Laboratories, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974-0636, U.S.A.

* Dept. of ECE, National Technical University of Athens, Zografou 15773, Athens, Greece.

Email: potam@research.bell-labs.com, maragos@cs.ntua.gr

May 13, 2008

Suggested keywords: *multiband demodulation, energy separation algorithm, AM–FM modulation model, pitch tracking, AM–FM vocoder, speech synthesis*

Pages: 28

Figures: 5

Tables: 1

Please address all correspondence to:

Alexandros Potamianos

Dialogue Systems Research

Bell Laboratories, Lucent Technologies

600 Mountain Ave., Room 2D-463

Murray Hill, NJ 07974-0636

U.S.A.

tel: (001) 908-582-4203

fax: (001) 908-582-7308

e-mail: potam@research.bell-labs.com

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

Abstract. In this paper, the AM–FM modulation model is applied to speech analysis, synthesis and coding. The AM–FM model represents the speech signal as the sum of formant resonance signals each of which contains amplitude and frequency modulation. Multiband filtering and demodulation using the energy separation algorithm are the basic tools used for speech analysis. First, multiband demodulation analysis (MDA) is applied to the problem of fundamental frequency estimation using the average instantaneous frequency as estimates of pitch harmonics. The MDA pitch tracking algorithm is shown to produce smooth and accurate fundamental frequency contours. Next, the AM–FM modulation vocoder is introduced, which represents speech as the sum of resonance signals. A time-varying filterbank is used to extract the formant bands and then the energy separation algorithm is used to demodulate the resonance signals into the 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 of speech resonances. Finally, the perceptual importance of modulations in speech resonances is investigated and it is shown that amplitude modulation patterns are both speaker and phone dependent.

Zusammenfassung. In diesem Artikel wird das AM–FM Modulationsmodel fuer Sprachanalyse, Sprachsynthese und Sprachkodierung angewendet. Das AM–FM Modulationsmodel repraesentiert das Sprachsignal als Summe von Formantresonanzen, welche jeweils Amplituden- und Frequenzmodulation enthalten. Multibandfilterung und Demodulation, basierend aus dem Energie-Trennungsalgorithmus, sind die wesentlichen Hilfsmittel fuer die Sprachanalyse. Zuerst wird die Multiband-Demodulationsanalyse (MDA), basierend auf der Schaetzung von Harmonischen der Grundfrequenz mittels durchschnittlicher Frequenz, auf das Problem der Grundfrequenzbestimmung angewandt. Der MDA-Algorithmus zur Bestimmung der Grundfrequenz erzeugt glatte und genaue Grundfrequenzverlauefe. Anschliessend wird der AM–FM-Modulationsvocoder, der Sprache als eine Summe von Resonanzsignalen darstellt, vorgestellt. Eine zeitlich variable Filterbank wird zur Extraktion der Formantbaender und der Energie-Trennungsalgorithmus zur Demodulation des Resonanzsignals in die Einhuellende der Amplitude und die zugrundeliegende Frequenz verwendet. Eine effiziente Modellierung und ein Kodierungsalgorithmus (4.8-9.6 kbits/sec) fuer die Einhuellende der Amplitude und die zugrundeliegende Frequenz der Sprachresonanzen werden vorgeschlagen. Abschliessend wird die perzeptuelle Bedeutung der Modulation von Sprachresonanzen untersucht und es wird gezeigt, dass die Muster der Amplitudenmodulation von Sprecher und Phonem abhaengen.

Résumé. Cet article présente un modèle de modulation AM–FM pour l'analyse, la synthèse, et le codage de la parole. Le modèle AM–FM décrit le signal de parole comme la somme de différents signaux représentant les fréquences formantiques, modulés en fréquence et amplitude. Un filtrage multibandes et une démodulation basée sur un algorithme de séparation d'énergie sont utilisés pour analyser le signal. Une analyse par démodulation multibandes (ADM) est tout d'abord employée afin d'estimer la fréquence fondamentale du signal, en se basant sur la fréquence instantanée moyenne comme estimation des hamoniques du pitch. Cet algorithme de suivi du pitch conduit à une estimation lisse et précise de la fréquence fondamentale. Un vocoder utilisant une modulation AM–FM est ensuite mis en oeuvre pour modéliser le signal par la somme de ses harmoniques. Un banc de filtres adaptatif permet d'extraire les bandes de fréquence formantiques et un algorithme fondé sur la séparation d'énergie est utilisé pour démoduler les harmoniques des formats en signaux instantanés modulés en amplitude et en fréquence. Différents algorithmes sont proposés conduisant à un codage efficace à 4.8-9.6 kbits/sec de l'enveloppe et la fréquence instantanée des résonances formantiques. Enfin, l'importance perceptive de la modulation des résonances du signal de parole est étudiée et démontre que la modulation d'amplitude ainsi obtenue est indépendante du locuteur et du phonème.

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 have had some success in reproducing and synthesizing speech using concatenative methods, but they have not been equally successful in transforming speaker characteristics and speaking styles in a controlled way.

Motivated by nonlinear and time-varying phenomena[1] during speech production and the need for a better understanding of the speech production process Maragos, Kaiser and Quatieri [14] proposed a nonlinear model that describes a speech resonance as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure

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

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 $f(t) = f_c + q(t)$. 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. Modeling formant resonance signals as AM–FM signals relates both to formant models and to the phase vocoder (see Section 3 for a comparison).

The use of a nonlinear model for speech resonances was motivated by the work of Teager [32]. In [11], Kaiser formally introduced the energy operator as a signal analysis tool. In a series of papers Maragos, Quatieri and Kaiser [16, 15, 14] laid down the groundwork for applying the AM–FM model and energy operators to demodulation of speech resonances. Multiband filtering and demodulation using the energy operator was formalized in [2]. Original work in the areas of signal processing, speech analysis, synthesis and recognition, music processing and image processing motivated or based on the AM–FM modulation model and the energy separation algorithm can be found in the literature [20, 13, 25, 31, 10, 8, 23]. Harmonic modeling using AM–FM models has also attracted some interest [24, 30] as a generalization of the the short-time invariant sinusoidal model [17]. Finally, the need of demodulating a sum of AM–FM signals with overlapping spectra

[1] Evidence for the existence of speech modulations has been provided in [14]. For instance, as Teager's experiments have demonstrated, the air jet flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses. This can cause instantaneous modulations of the amplitude and frequency of a speech resonance as explained in [14] using time-varying oscillators.

was the motivation for the work in [28].

In this paper, the AM–FM modulation model is applied to speech analysis, synthesis and coding. The first part of this paper continues the work of [22] on multiband demodulation for speech analysis applications. Specifically, multiband demodulation is applied to fundamental frequency estimation: the average instantaneous frequency in each band is used as a harmonic frequency estimate (i.e., fundamental frequency multiple) and the fundamental frequency is calculated from the harmonic frequencies by functional minimization. The second part of this paper focuses on the AM–FM modulation vocoder as a means to model and study the perceptual importance of modulations in speech. The vocoder models speech as a sum of formant resonance signals (see Eq. (1)) extracted from the speech signal through time-varying filtering. Efficient algorithms are proposed for modeling and coding of the amplitude envelope and instantaneous frequency signals of each resonance. The modeling provides flexibility in controlling the amount of amplitude and frequency modulations in the synthesized resonance signals. The proposed analysis-synthesis system provides the means for measuring the amount and perceptual importance of amplitude and frequency modulation in speech resonances. Overall, the paper offers a comprehensive collection of algorithms that can be used for the analysis and synthesis of nonlinear phenomena in speech production.

The organization of this paper is as follows: First multiband demodulation is introduced, the analysis tool used extensively in this paper. In Section 2.4 the 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.

2 Speech Analysis

In this section, the main tools used throughout the paper are introduced namely multiband filtering and demodulation analysis. Short-time instantaneous frequency estimates are proposed and their relative merits are discussed for formant and fundamental frequency estimation. Next, the multiband demodulation formant tracking algorithm introduced in [22] is outlined. Finally, multiband filtering and demodulation are applied to the problem of fundamental frequency estimation.

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 Gabor filter, by being maximally smooth and optimally concentrated both in the time and frequency domain, provides smooth amplitude and frequency estimates in the demodulation stage that follows. The amplitude envelope $|a(t)|$ and 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) [14] is based on the nonlinear differential Teager–Kaiser energy operator [11]. 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 [14]

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

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 [21] for a ESA vs. HTD comparison).

2.2 Short-time Frequency Estimates

Short-time estimates of the average instantaneous frequency have been proposed in [22] in the context of an MDA-based formant tracking application. Specifically, the unweighted F_u and weighted average instantaneous frequency F_w estimates were defined as (see also [4])

$$F_u = \tfrac{1}{T} \int_{t_0}^{t_0+T} f(t)\, dt \tag{3}$$

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)\, [a(t)]^2\, dt}{\int_{t_0}^{t_0+T} [a(t)]^2\, dt} \tag{4}$$

where $|a(t)|$, $f(t)$ are the amplitude envelope and the instantaneous frequency signals of resonance signal $r(t)$, t_0 and T are the start and duration of the analysis frame.

To illustrate the behavior of the two short-time estimators we assume that a speech signal can be modeled as a sum of sinusoids with slowly time-varying amplitudes and frequencies [17] in particular, a speech resonance can be modeled as a sum of a few sinusoids representing the

harmonics in the formant band, i.e.,

$$r(t) = \sum_n a_n \cos[2\pi f_n t + \theta_n] \tag{5}$$

where a_n, f_n, θ_n are the harmonic amplitudes, frequencies and phases. Using this simple sinusoidal model it can be shown (see Appendix A) that F_u locks on the harmonic frequency with the greatest amplitude in the formant band, while F_w weights each harmonic frequency in the formant band by its squared amplitude, i.e.,

$$F_u \approx f_M, \qquad F_w = \frac{\sum_n f_n a_n^2}{\sum_n a_n^2} \tag{6}$$

where f_M is the frequency of the most prominent harmonic in the spectrum band ($a_M = \max_k(a_k)$). As a result, the unweighted estimate F_u is a good harmonic frequency estimate, i.e., a multiple $k\,\mathrm{F0}$ of the fundamental frequency $F0$. The accuracy of the harmonic estimate improves as the bandwidth of the resonance signal $r(t)$ decreases. The weighted estimate F_w provides an amplitude weighted harmonic frequency average which is a natural formant frequency estimate. For that reason F_w was used for formant tracking in [22]. A visualization of the properties of F_u, F_w for the sum of two amplitude modulated sinusoids can be found in [22].

2.3 Formant Tracking

In [22], multiband demodulation analysis and the weighted frequency estimate F_w were applied to formant tracking. The MDA formant tracking algorithm is briefly reviewed 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). 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 (every 10 ms) for each speech frame located around time t 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. For a detailed explanation of the MDA formant tracker, results and comparisons with other formant tracking algorithms see [22].

2.4 Fundamental Frequency Estimation

As discussed in Section 2.2 and shown in the appendix, 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 $r(t)$). Next, a multiband demodulation pitch tracking algorithm is proposed using F_u as a harmonic frequency estimate[2].

Similarly to MDA formant tracking, the speech signal is filtered through a bank of Gabor bandpass filters and then each filtered signal is demodulated to amplitude envelope and instantaneous frequency signals. Typical effective RMS Gabor filter bandwidth is 200 Hz and the approximate spacing is 100 Hz following a mel frequency scale (by using a non-uniform filter spacing harmonic frequency estimation errors are averaged out). 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 resulting time-frequency average instantaneous frequency distribution $F_u(t,\nu)$ is used in a functional minimization procedure to estimate the pitch contour. Fig. 1

A typical Gabor filterbank is shown in Fig. 1(a). The harmonic frequency estimates for a 20 ms speech frame are shown as dotted lines superimposed on the Fourier spectrum of the speech signal in Fig. 1(b). Note that certain harmonics have no corresponding F_u estimates while others have more than one estimates depending on the position of the filters, i.e., MDA is a non-parametric analysis method. The time-frequency distribution of $F_u(t,\nu)$ is shown in Fig. 2(c) for a sentence from the TIMIT database. The harmonic tracks are clearly visible and directly correspond to the harmonic regions in the narrowband speech spectrogram shown in Fig. 2(b). Fig. 2

The fundamental frequency of a voiced speech segment is determined from the minimization of the weighted error sum $E(\mathrm{F0})$ over all possible fundamental frequency candidates F0

$$E(\mathrm{F0}) = \frac{1}{\mathrm{F0}} \sum_{n=1}^{N} \alpha(\nu_n) \mid F_u(\nu_n) - \lfloor \frac{F_u(\nu_n)}{\mathrm{F0}} + 0.5 \rfloor \ \mathrm{F0} \mid \qquad (7)$$

where $\lfloor . \rfloor$ denotes truncation of the decimal part and $\lfloor . + 0.5 \rfloor$ is the rounding operator, 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. The weighting factors $\alpha(\nu_n) =< a^2(t,\nu_n) >_T$ measure the relative prominence of the estimated harmonic $F_u(\nu_n)$. In the error sum of Eq. (7), deviations of the harmonic 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 $F_u(\nu_n), n = 1, 2..$ and the fundamental

[2] Alternatively, the slope of the phase signal $S_\phi = \frac{1}{2\pi} \frac{\int_{t_0}^{t_0+T} t\,\phi(t)\,dt}{\int_{t_0}^{t_0+T} t^2\,dt}$ computed from linear regression can provide more noise-robust estimates (the phase signal is the integral of the instantaneous frequency signal: $\phi(t) = 2\pi \int_{-\infty}^{t} f(\tau)\,d\tau$)[20].

frequency multiples k F0, $k = 1, 2...$ The algorithm produces very detailed and smooth fundamental frequency contours as shown in the example of Fig. 2(d) for the speech signal in Fig. 2(a).

The pitch contours are filtered by a median filter 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 F0(t) is defined as

$$E_G = \int_{t_1}^{t_2} E[\mathrm{F0}(t)]\, dt + \lambda \int_{t_1}^{t_2} \left(\frac{d\mathrm{F0}(t)}{dt} \right)^2 dt \tag{8}$$

for each voiced region $[t_1, t_2]$. E is the error criterion of Eq. (7) 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 $\ll 1$ Hz) with a small 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, it is shown in the appendix that when the analysis window duration T is a multiple of the pitch period the accuracy of the F_u estimate is

$$F_u = f_M + O(\epsilon^4)\ , \quad \epsilon = \max_{k \neq M}(a_k / a_M) \tag{9}$$

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 arbitrary window duration T. Pitch-synchronous refinement of fundamental frequency contours was shown to eliminate pitch estimation errors for synthetic speech signals.

The MDA pitch tracker is related to pitch trackers based on the sinusoidal model [17]. Both algorithms estimate the most prominent harmonics in the speech spectrum and use a functional minimization approach to determine the pitch contour [18, 7]. The MDA pitch tracker is also related to auditory filterbank processing. In [19], McEachern speculates that the fundamental frequency is perceived as a weighted sum of the harmonic frequencies estimated for each auditory filter through demodulation. In [27], Quatieri et al propose perceptually-motivated demodulation algorithms that use the output of two filters with overlapping frequency responses.

The pitch tracker was evaluated on 37 utterances from the TIMIT database. Each sentence was spoken from a different speaker (23 male, 14 female speakers). The MDA pitch tracker filterbank (as tested) consisted of 20 mel-spaced filters spanning the 0-2000 Hz range. The MDA pitch estimates were compared to the pitch estimates computed by the ESPS signal processing package (of Entropic Research Laboratory) based on [29]. A 40 ms analysis window (updated every 10 ms) was used for both pitch trackers. The following were the main results from visual inspection of the tracks and

from detailed numerical comparisons:

- The tracks of the ESPS tracker were over-smoothed, especially in the voiced-unvoiced transition regions.
- The total number of segments where the estimated pitch was approximately half or double the actual value were twelve for the ESPS tracker vs. five for the MDA pitch tracker (a segment includes at least four consecutive frames with halving or doubling errors).
- The mean and standard deviation of the differences in fundamental frequency estimates between the ESPS and MDA pitch trackers (pitch doublings and halvings excluded) computed over all sentences was 0.7 Hz (ESPS minus MDA) and 3.1 Hz respectively. For each sentence, the range for the mean difference was 0 to 1.9 Hz and for the standard deviation 1.6 to 5.9 Hz.

Overall, the multiband filtering and demodulation pitch tracking algorithm is simple, and produces smooth and accurate fundamental frequency contours.

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 multiband demodulation formant tracking algorithm (see Section 2.3). 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. **??**.

Both the AM–FM vocoder and the the parallel formant vocoder [9, 12] 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 coupling between the excitation and vocal tract, the AM–FM modulation model allows us more freedom to investigate nonlinear speech production phenomena not modeled by the source-linear filter model. The representation of a speech band by the amplitude envelope and instantaneous frequency signals is common ground between the AM–FM vocoder and the phase vocoder [5, 6]. The main difference is that the AM–FM vocoder uses a time-varying filterbank to extract the formant bands, while the phase vocoder uses a bank of filters fixed in frequency. In addition, most implementations of the phase vocoder use narrow frequency bands that span one or two harmonics, while each frequency band of the AM–FM vocoder contains a formant spectral peak that typically comprises of six to seven harmonics. As a result, the structure of the information signals is also different, and novel algorithms have to be devised to efficiently capture the patterns in the amplitude envelope and instantaneous frequency signals of the AM–FM vocoder. In the next section, efficient modeling and coding algorithms for the amplitude envelope and instantaneous frequency signals of speech resonances are proposed.

3.1 Modeling 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

$$a(n) = u(n) * g(n) * h(n), \quad u(n) = \sum_{k=1}^{K} b_k \delta(n - n_k) \tag{10}$$

where the impulse sequence $u(n)$ is the excitation signal, $g(n)$ is the impulse response of a critically damped second-order system and $h(n)$ is the baseband impulse response of the filter used for extracting the corresponding resonance signal $r(t)$ (for a real Gabor filter $h(t) = exp(-\alpha t^2)$). The frequency response $G(z)$ of the critically damped system with impulse response $g(n)$ is

$$G(z) = c_0/1 + c_1 z^{-1} + c_2 z^{-2}, \; c_1 = -2e^{-\pi B/F_s}, \; c_2 = e^{-2\pi B/F_s} \tag{11}$$

where B determines the rate of decay of the amplitude envelope signal and F_s is the sampling frequency. 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 $a(n)$. The impulse response of this one-parameter 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, $h(n)$ was introduced in the amplitude envelope model of Eq. (10) 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, 20] so that the mean square modeling error

$$E = \sum_{n=1}^{N} e(n)^2 = \sum_{n=1}^{N} [s(n) - \hat{s}(n)]^2 \tag{12}$$

is minimized, where N is the size of the speech analysis frame. The analysis-by-synthesis loop is shown in Fig. 3. Note that the use of a second order linear model for the amplitude envelope is only meant as a simple mathematical parameterization of the amplitude and is not otherwise related to the physics of speech production. The model can efficiently capture amplitude modulation patterns and offers control over the amplitude modulation amount in speech resonances. For example, the average signal to noise ratio for the modeled amplitude envelope signals was 20 dB when using about two pulses per pitch period. Five sentences from the TIMIT database were used for the test (same as in Section 3.2).

In Fig. 4(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. 4(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. Fig. 4

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 is allowed to take different values for the open and closed phase of voicing. Note that such frequency modulation patterns will cause corresponding bandwidth modulations (according to the governing equations of the linear second-order oscillator) that have to be accounted for as changes in the rate of decay of the corresponding amplitude envelope signal. In Fig. 4(c), the (actual) instantaneous frequency signals and formant tracks (dashed) are shown

for F1 and F2.

3.2 Coding 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. Primary pulses are typically located close to the major excitation instants (one per pitch period for voiced speech), while secondary pulses model secondary excitations and nonlinear production phenomena. Pulses are labeled as primary or secondary as follows: (a) regions of steady-state voicing (probability of voicing ≈ 1) are identified, (b) for steady-state voicing regions the primary pulses are labeled based on their greater amplitudes and periodic spacing, (c) starting from these voiced "anchor" segments the neighboring primary pulses are determined by searching for the pulse with the greatest amplitude in a window centered one pitch period apart from the current primary pulse; this process is repeated both forward and backward in time until the signal support is covered, (d) the remaining pulses are classified as secondary. The distances between consecutive primary pulses form a slowly time-varying contour ("pitch" contour) which can be efficiently quantized. Similarly, the amplitudes of the excitation pulses form a "smooth" contour and are quantized using PCM. Secondary pulse positions are coded relatively to the primary ones. Finally, the envelope decay rate parameter B is sampled at 100 Hz and quantized using PCM. Typically 3.5 to 6 kbits/sec are used to quantize the envelope signals with good detail. The amplitude envelope signals are reconstructed from the excitation signals $u(n)$ using Eq. (10), where $g(n)$ is computed using the quantized B values and $h(n)$ is determined from the bandwidths of the analysis filterbank.

As discussed in the previous section the instantaneous frequency signals have two major components: the average formant frequencies and frequency modulation around the formant tracks. The average formant frequencies are computed as the amplitude weighted instantaneous frequency average (F_w estimate). Short-time deviations from the formant values are measured for the open and closed excitation phase. Specifically, F_w is computed separately for speech segments that lay between primary and secondary excitation locations (roughly corresponding to the closed excitation phase for voiced speech) and for segments between secondary and primary excitation locations (open phase). Thus, a piece-wise linear model is assumed for the instantaneous frequency signals. The slow- and fast-varying components of the instantaneous frequency signals are coded separately. Formant tracks are decimated to 60 Hz and quantized using PCM. The frequency modulation components for the open and closed phase are coded separately as deviations from the average formant

frequency (only for the first formant where FM is perceptually most important). Finally, the absolute phase at primary excitation instants was judged to be perceptually important for formant bands below 1000 Hz. Typically 1.3 to 3 kbits/sec are alloted to the instantaneous frequency signals. The instantaneous frequency signals are reconstructed by interpolation of the quantized formant frequency tracks. Then the FM component is added and the phase signal is obtained as the running integral of the instantaneous frequency signal. Finally, a phase discontinuity is added at envelope minima to guarantee that the phase at excitation instants takes the appropriate value. A moving average filter is used to smooth the discontinuity[3].

The resonance signals are reconstructed from the amplitude envelope and phase signals as in Eq. (1), and added to obtain the coded speech signal. A typical bit allocation scheme for the various components of the AM–FM modulation vocoder is shown in Table 1. Table 1

Informal listening tests were performed for both the analysis-synthesis system and the vocoder (at various bit rates) on sentences of the TIMIT database spoken by different speakers. A comparative test was performed among five non-expert listeners to evaluate the speech quality at various stages of the modeling and coding process. Each listener was presented with eight (arbitrarily chosen) TIMIT sentences. The listeners were given the original signal (bandpassed between 200 and 5000 Hz) and then presented with the following signals (in arbitrary order): (1) the sum of resonance bands, (2) the coded signal (not quantized), (3) the coded and quantized signal at 4.8 kbits/sec, (4) the DoD-LPC coded signal at 2.4 kbits/sec [33] and (5) the DoD-CELP coded signal at 4.8 kbits/sec [3]. The listeners rated the quality of each of the five signals from one to five (best). The results were as follows (the mean and variance of the ratings were normalized for each speaker): (1) 4.1, (2) 3.6, (3) 2.5, (4) 3.2, and (5) 4.0. Speech coded (but not quantized) by the AM–FM vocoder was rated between the LPC and CELP vocoders, while quantized speech was rated below both reference vocoders. The low quality of the quantized speech is mostly due to phase quantization artifacts in the low frequency bands that disappear at higher bit rates. More work is needed to improve the quantization algorithms and to produce high quality speech around 4.8 kbits/sec. The AM–FM analysis-synthesis system (no coding or quantization) scores the highest out of all systems.

[3]The instantaneous phase signal is discontinuous at excitation instants, or equivalently the instantaneous frequency signal (being the derivative of the phase) presents a single or double spike. When estimating the instantaneous frequency the discontinuity is smoothed as a side-effect of band-pass filtering. This effect is reproduced when coding and quantizing the instantaneous frequency signal (see also [20]).

Overall, the results are encouraging and show that more work is needed to optimize the quantization algorithms, to improve modeling and to test the vocoder under adverse conditions. Specifically, the modeling of the very low (0-200 Hz) and very high (> 5 kHz) frequency regions is inadequate. Further, spectral zeros are not modeled. Additional work is needed to improve the efficiency of the coding and quantization algorithms especially for the instantaneous frequency signals in order to produce high-quality coded speech at 4.8 to 9.6 kbits/sec. However, even with the current simple implementation, the AM–FM analysis-synthesis system produces natural speech and provides the test-bed for the perceptual importance of modulations in speech.

4 Discussion

In this section, the perceptual importance of amplitude and frequency modulations is discussed. First we present preliminary results that show that modulation patterns are both speaker and phone dependent and could provide important perceptual cues for noise-corrupted or bandpassed speech. Further, alternative ways of modeling the amplitude and frequency modulation patterns are proposed.

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 (see also next paragraph). 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 speaker and phone dependency of amplitude modulation patterns was verified by conducting a preliminary analysis of 120 sentences of the TIMIT database collected from 12 male speakers (10 sentences per speaker) using the AM–FM modulation vocoder. Each sentence was analyzed using the techniques outlined in the previous sections. For each sentence in the database, the primary and secondary pulse locations and amplitudes were computed for each formant resonance amplitude envelope signal (F1, F2 and F3). Next the average ratio of the secondary to primary excitation pulse amplitude was computed as a rough estimate of the amplitude modulation in-

dex. The modulation index estimate was computed for 15 monophtongal vowels and diphthongs (using the phonemic segmentation and labels provided with the TIMIT database). Average modulation index estimates were computed for each *phone*, for each *speaker*, and for each left and right phonemic group *context*. It was found that the amount of modulation in each band was speaker-dependent ranging: 13-24% (F1), 14-40% (F2) and 9-40% (F3). Average AM index values for all 120 sentences analyzed were 16% (F1), 23% (F2), 23% (F3). The range of phone-dependent AM index estimates was 13-19% (F1), 17-30% (F2) and 17-30% (F3). Phonemes that displayed the highest amount of modulation were /ao/, /ax/, while /aw/, /eh/ displayed the lowest. Finally, context-dependent AM modulation indexes were computed. The following left and right contexts were investigated: silence, vowel, plosive, nasal, glide, voiced fricative and unvoiced fricative. The AM index was found to be 30% higher than average for segments preceded or followed by silence. Increased AM amounts were also found in the context of glides and voiced fricatives. This is to be expected since the dynamics of speech production are changing rapidly during voiced-unvoiced transitions, silence-speech transitions, and glides. Further investigation is required to understand the importance of the modulation patterns for speech synthesis and recognition.

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 [17] 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 [26]. 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.

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.

Overall, the AM–FM analysis-synthesis system accounts for a variety of speech production phenomena not described in linear models and, as a result, produces speech of 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.

Appendix A

Consider the sum of N sinusoids with constant[4] amplitudes a_n and frequencies f_n

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

where θ_n are arbitrary phase constants. Assuming that the bandwidth of $r(t)$ is much smaller than $\min_n(f_n)$, the analytical signal $z(t)$ estimates for the amplitude envelope $|a_H(t)|$ and instantaneous frequency $f_H(t)$ computed from the Hilbert transform are

$$|a_H(t)| = |z(t)| \approx (\; \textstyle\sum_n \sum_k \; a_n a_k \cos[\Delta\phi_{nk}(t)] \;)^{\frac{1}{2}} \tag{A2}$$

$$f_H(t) = \tfrac{d}{dt}\angle z(t) \approx \textstyle\sum_n \sum_k \; f_n a_n a_k \cos[\Delta\phi_{nk}(t)]/[a_H(t)]^2 \tag{A3}$$

where $\Delta\phi_{nk}(t) = 2\pi[f_n - f_k]t + (\theta_n - \theta_k)$.

We will show next that under the assumption $|f_{i+1} - f_i| = F_0$, $i = 1...N-1$

$$F_u = \frac{1}{T} \int_t^{t+T} f_H(t)\; dt \approx f_m, \quad \text{if} \quad a_m >> a_1, ..., a_{m-1}, a_{m+1}...a_N \tag{A4}$$

i.e., for a sum of harmonically related sinusoids the unweighted instantaneous frequency F_u locks onto the frequency of the sinusoid with the greatest amplitude. By expanding the denominator in Eq. (A3) in a Taylor series

$$f_H(t) \approx \Big(f_m + \sum_{n\neq m} \frac{a_n}{a_m}\; (f_n + f_m)\; \cos[\Delta\phi_{nm}] + \sum_{n\neq m} (\frac{a_n}{a_m})^2\; f_n + \sum_{n\neq m} \sum_{k\neq m,n} \frac{a_n\; a_k}{a_m^2}\; f_k\; \cos[\Delta\phi_{nk}]\Big)$$

$$\Big(1 - 2\sum_{n\neq m} \frac{a_n}{a_m}\; \cos[\Delta\phi_{nm}] - \sum_{n\neq m} (\frac{a_n}{a_m})^2 - \sum_{n\neq m} \sum_{k\neq m,n} \frac{a_n\; a_k}{a_m^2}\; \cos[\Delta\phi_{nk}] + 4\sum_{n\neq m} (\frac{a_n}{a_m})^2\; \cos^2[\Delta\phi_{nm}] + \text{h.o.t.}\Big)$$

We further assume that the averaging window duration T is proportional to the "pitch period" of the sum of the sinusoids, i.e., $T \propto 1/F_0$ (pitch-synchronous analysis). In this case $\int_t^{t+T} \cos[\Delta\phi_{nk}]\; dt = 0$ and $\int_t^{t+T} \cos[\Delta\phi_{nk}]\; \cos[\Delta\phi_{ij}]\; dt = 0$, for $(n,k) \neq (i,j)$. Carrying out the algebra

$$\begin{aligned} F_u &= \tfrac{1}{T} \int_t^{t+T} f_H(t)\; dt \\ &\approx f_m - \textstyle\sum_{n\neq m} (\frac{a_n}{a_m})^2\; f_m + 2\sum_{n\neq m} (\frac{a_n}{a_m})^2\; f_m - \sum_{n\neq m} (\frac{a_n}{a_m})^2\; (f_n + f_m) + \sum_{n\neq m} (\frac{a_n}{a_m})^2\; f_n + \text{h.o.t.} \\ &= f_m + O(\epsilon^4)\;, \quad a_m >> a_1, ..., a_{m-1}, a_{m+1}...a_N \end{aligned}$$

[4]Note that the results presented in the appendix hold approximately for sums of sinusoids with time-varying amplitudes $a_n(t)$ and frequencies $f_n(t)$, provided that $a_n(t)$ and $f_n(t)$ are slowly-varying compared to $\cos[2\pi f_n t]$. In this case, $a_n(t)$ and $f_n(t)$ can be assumed constant when differentiating or integrating in the presence of the fast varying $\cos[2\pi f_n t]$ term (two time-scale analysis).

since the $O(\epsilon^2)$ terms cancel out, where $\epsilon = \max_{k \neq m}(a_k/a_m)$. Thus, for pitch-synchronous analysis, the approximation error is $O(\epsilon^4)$. In practice, the duration of the averaging window used is not a multiple of the "pitch period" and the exact values of F_u depend on the averaging window boundaries. For pitch-asynchronous analysis, the order of the approximation error is $O(\epsilon)$.

Similarly, for the weighted estimator F_w one may write

$$F_w = \frac{\int_t^{t+T} f_H(t)[a_H(t)]^2 \, dt}{\int_t^{t+T} [a_H(t)]^2 \, dt} = \frac{\sum_n \sum_k \; f_n a_n a_k \int_t^{t+T} \cos[\Delta\phi_{nk}(t)] \, dt}{\sum_n \sum_k \; a_n a_k \int_t^{t+T} \cos[\Delta\phi_{nk}(t)] \, dt} = \frac{\sum_n f_n a_n^2}{\sum_n a_n^2} \qquad \text{(A5)}$$

i.e., the weighted instantaneous frequency F_w equals the amplitude weighted average of the hamornic frequencies.

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. Internat. Conf. on Acoust., Speech, and Signal Process.*, (Paris, France), pp. 614–617, May 1982.

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

[3] J. P. Campbell, T. E. Tremain and V. C. Welch, "The Federal Standard 1016 4800 bps CELP Voice Coder," *Digital Signal Processing, Academic Press*, vol. 1, no. 3, pp. 145–155.

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

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

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

[7] E. B. George, *An Analysis–by–Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Signal Processing.* Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, 1991.

[8] J. P. Havlicek, *AM–FM Image Models.* Ph.D. Thesis, University of Texas at Austin, Austin, TX, 1996.

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

[10] C. R. Jankowski, *Fine Structure Features for Speaker Identification.* Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996.

[11] Kaiser, J. F.,"On Teager's Energy Algorithm and Its Generalization to Continuous Signals," in *Proc. 4th IEEE DSP Workshop*, Mohonk, New Paltz, NY, Sep. 1990.

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

[13] S. Lu and P. C. Doerschuk, "Nonlinear Modeling and Processing of Speech Based on Sums of AM–FM Formant Models", *IEEE Trans. Signal Processing*, vol. 44, no. 4, pp. 773–782, Apr. 1996.

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

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

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

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

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

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

[20] A. Potamianos, *Speech processing applications using an AM–FM modulation model.* Ph.D. Thesis, Harvard University, Cambridge, MA, 1995.

[21] 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.

[22] 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

[23] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," in *Proc. European Conf. on Speech Communications and Technology*, (Rhodes, Greece), pp. 1355–1358, Sept. 1997.

[24] M. A. Ramalho, *The Pitch Mode Modulation Model with Applications in Speech Processing*. Ph.D. Thesis, The State University of New Jersey, New Brunswick, NJ, Jan. 1994.

[25] P. Rao, "A Robust Method for the Estimation of Formant Frequency Modulation in Speech Signals", *Proc. IEEE ICASSP–96*, Atlanta, Georgia, pp. II–813–816, May 7–10, 1996.

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

[27] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM–FM separations using auditory-motivated filters," *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 465–480, 1997.

[28] B. Santhanam, *Multicomponent AM–FM Energy Demodulation with Application to Signal Processing and Communications*. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, 1998.

[29] B. G. Secrest, G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, pp. 1352-1355, 1983.

[30] Y. Stylianou, "Decomposition of Speech Signals into a Deterministic and a Stochastic Part," in *Proc. Internat. Conf. on Spoken Lang. Process.*, pp. 1213-1216, Philadelphia, PA, 1996.

[31] R. B. Sussman, *Analysis and Resynthesis of Musical Instrument Sounds Using Energy Separation*, Master's Thesis, Rutgers, The State University of New Jersey, Graduate School, New Brunswick, NJ, Apr. 1996.

[32] H. M. Teager and S. M. Teager, "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, 1990.

[33] T. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology*, pp. 40-49, July 1982.

List of Footnotes

1. Evidence for the existence of speech modulations has been provided in [14]. For instance, as Teager's experiments have demonstrated, the air jet flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses. This can cause instantaneous modulations of the amplitude and frequency of a speech resonance as explained in [14] using time-varying oscillators.

2. Alternatively, the slope of the phase signal $S_\phi = \frac{1}{2\pi} \frac{\int_{t_0}^{t_0+T} t\,\phi(t)\,dt}{\int_{t_0}^{t_0+T} t^2\,dt}$ computed from linear regression can provide more noise-robust estimates (the phase signal is the integral of the instantaneous frequency signal: $\phi(t) = 2\pi \int_{-\infty}^{t} f(\tau)\,d\tau$)[20].

3. The instantaneous phase signal is discontinuous at excitation instants, or equivalently the instantaneous frequency signal (being the derivative of the phase) presents a single or double spike. When estimating the instantaneous frequency the discontinuity is smoothed as a side-effect of band-pass filtering. This effect is reproduced when coding and quantizing the instantaneous frequency signal (see also [20]).

4. Note that the results presented in the appendix hold approximately for sums of sinusoids with time-varying amplitudes $a_n(t)$ and frequencies $f_n(t)$, provided that $a_n(t)$ and $f_n(t)$ are slowly-varying compared to $\cos[2\pi f_n t]$. In this case, $a_n(t)$ and $f_n(t)$ can be assumed constant when differentiating or integrating in the presence of the fast varying $\cos[2\pi f_n t]$ term (two time-scale analysis).

List of Figures

List of Tables

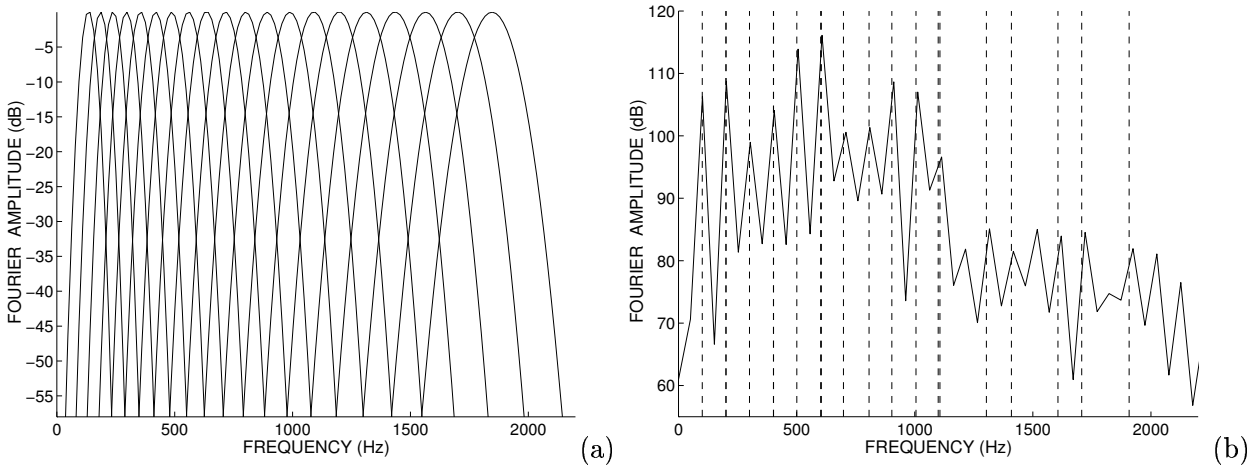

Figure 1: (a) The Mel–spaced "dense" Gabor filterbank used for MDA and (b) the average instantaneous frequency $F_u(\nu)$ estimates for each frequency band ν shown superimposed on the Fourier spectrum for a 20 ms speech frame (/aa/ from "dog").

Information source				bits/sec
amplitude envelope	envelope excitation	primary excitation	position	550
			amplitude	1400
		secondary excitation	position	275
			amplitude	700
	envelope rate of decay B			200
instantaneous frequency	average formant frequency			1270
	frequency modulation (F1 only)			240
	phase at primary excitation (F1 only)			260
Total (bps)				4895

Table 1: A typical bit allocation scheme for the AM–FM modulation vocoder.

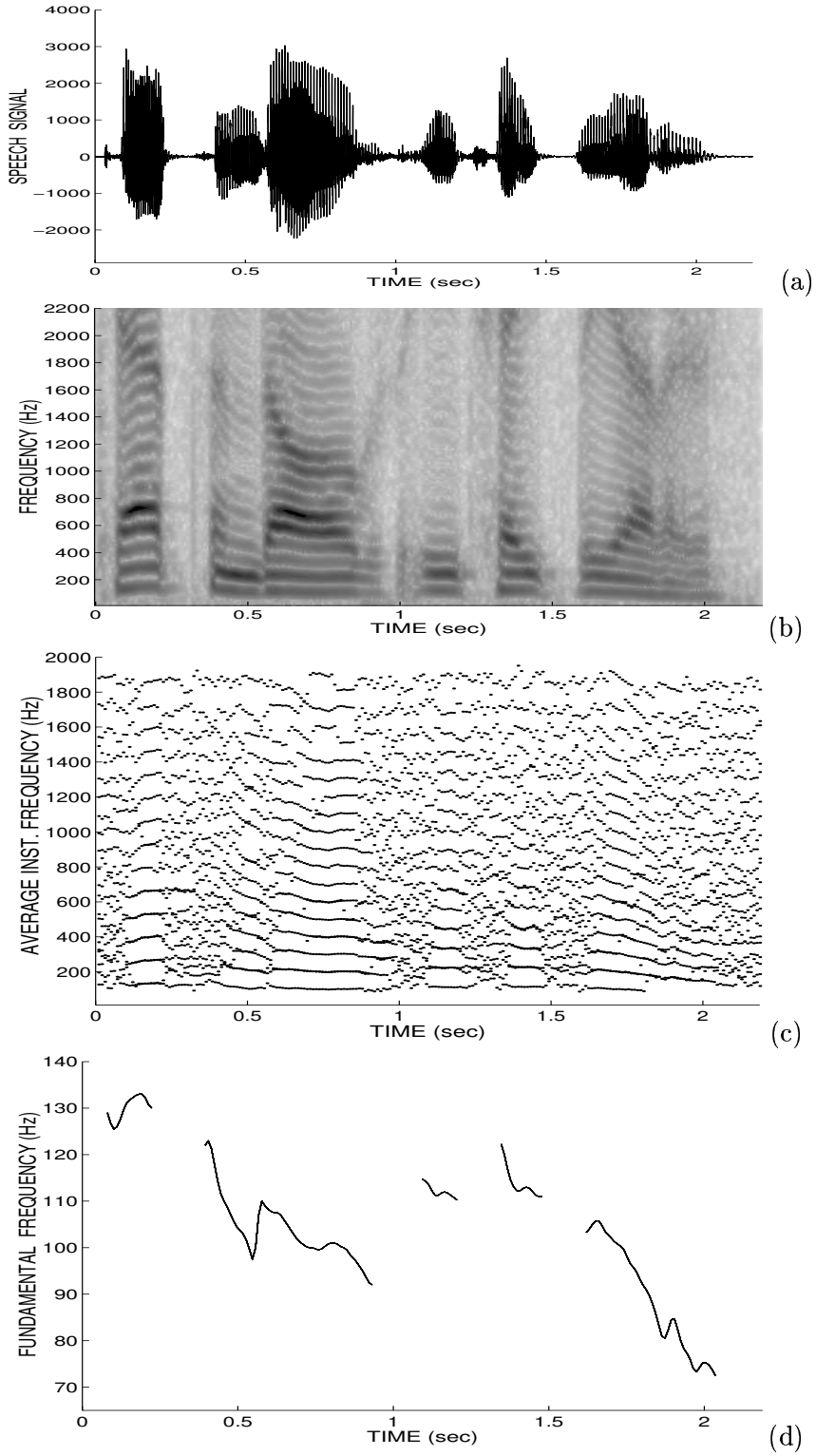

Figure 2: (a) Speech signal: "Cats and dogs each hate the other." (b) Narrowband speech spectrogram. (c) Time-Frequency average instantaneous frequency distribution (20 ms window). (d) MDA fundamental frequency contour.

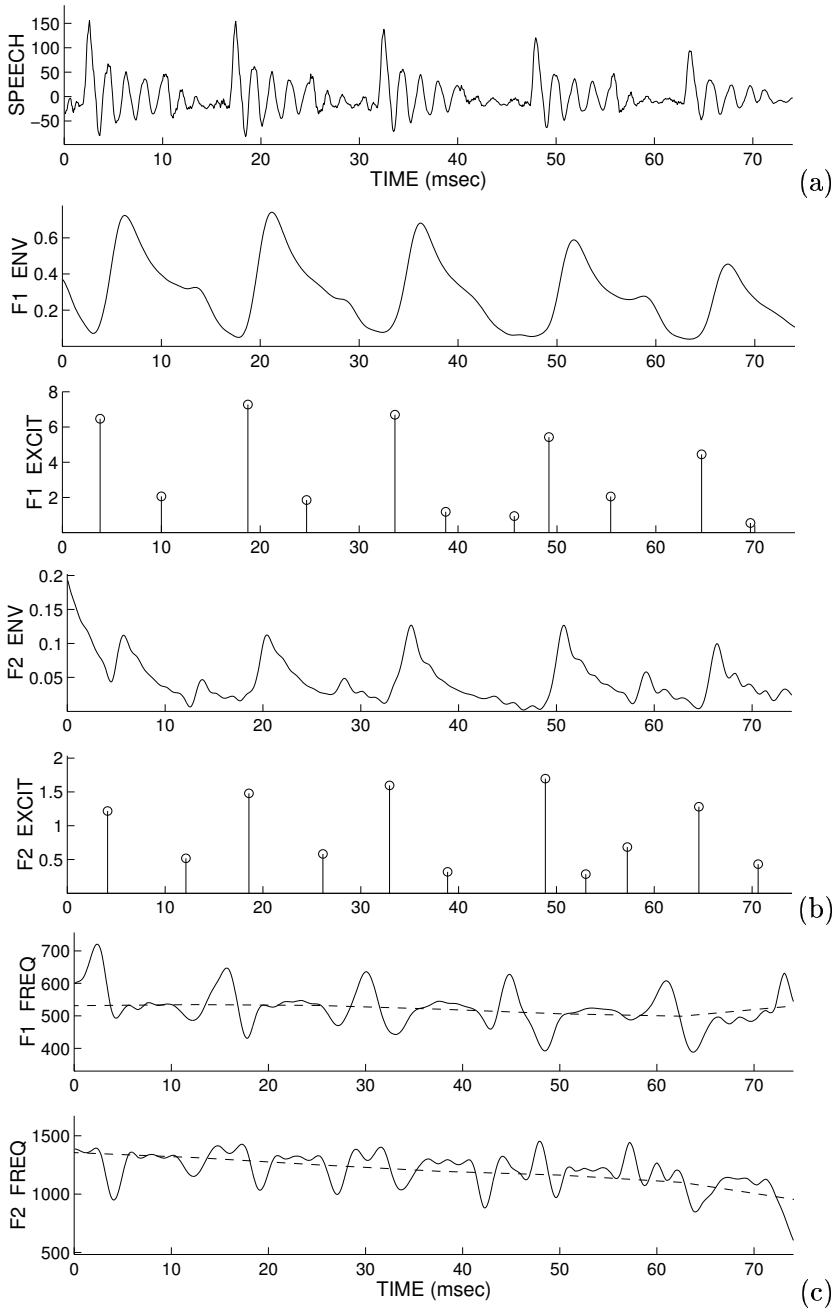

Figure 3: (a) Speech signal, phoneme /ow/ from "zero". (b) Amplitude envelope and multipulse excitation signals for the first and second resonances. (c) Instantaneous frequency signals and formant tracks for F1, F2.