FINDING SPEECH FORMANTS AND MODULATIONS VIA ENERGY SEPARATION: WITH APPLICATION TO A VOCODER

Helen M. Hanson *Petros Maragos* *Alexandros Potamianos*
Division of Applied Sciences, Harvard University, Cambridge, MA 02138, USA

ABSTRACT

An experimental system that uses an energy-tracking operator and a related energy separation algorithm to automatically find speech formants and amplitude/frequency modulations in voiced speech segments is presented. Initial estimates of formant center frequencies are provided by either LPC or morphological spectral peak picking. These estimates are improved by a combination of bandpass filtering and iterative application of energy separation. The system is shown to be effective. Its application to an AM-FM vocoder is also discussed.

1. INTRODUCTION

The ability to automatically find and track resonant frequencies of the speech production system, called 'formants', is an important part of speech processing, because formants play a major role in most speech applications. Traditional methods for formant finding are peak picking of the cepstrally-smoothed or LPC spectrum, or finding the roots of the LPC polynomial. These methods assume that the formants are constant within an analysis frame. However, in a recently proposed modulation model [1, 2], resonances are modeled as damped AM-FM signals

$$\begin{aligned} x(t) &= a(t)\cos[\phi(t)] \\ &= a(t)\cos[2\pi(f_c t + \int_0^t q(\tau)d\tau) + \phi(0)] \qquad (1) \end{aligned}$$

with a time-varying instantaneous frequency (in Hz)

$$f(t) = \frac{1}{2\pi}\dot{\phi}(t) = f_c + q(t) \qquad (2)$$

and a time-varying (generally non-exponential) amplitude $a(t)$. To estimate the amplitude and frequency signals $a(t), f(t)$, Maragos, Kaiser, and Quatieri [2] have developed an *energy separation algorithm (ESA)* that uses a nonlinear energy operator to track the instantaneous energy of the source generating the AM–FM signal and separate it into its amplitude and frequency components. This *energy operator*, defined as

$$\Psi_c[x(t)] = (\dot{x}(t))^2 - x(t)\ddot{x}(t) \qquad (3)$$

was introduced by Teager and Kaiser [3].

The ESA, however, is applied to single speech resonances, while the speech signal itself is multi-component, being the sum of several resonances. Thus, there is a need to isolate resonances by bandpass filtering. In our work this is done with a Gabor filter whose impulse response is $g(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t)$, $\omega_c = 2\pi f_c$. There is a question as to the best method of choosing the center frequency f_c and bandwidth parameter α of the filter. The carrier frequency f_c of the AM-FM signal may be a logical choice, but determining that frequency from an arbitrary signal may not be straightforward. Furthermore, the choice of filter bandwidth is complicated by conflicting requirements: the filter must be as wide as is possible to include the desired formant modulations, but narrow enough to exclude those of neighboring formants.

We would like to automatically determine the center frequencies of the bandpass filters used to extract the component AM-FM signals of the speech segment and then determine the modulations around these center frequencies, assuming we have a reasonable estimate of the filter bandwidths. The system presented in this paper is aimed at achieving this goal. The outline of the paper is as follows. We first review the energy operator and ESA in discrete time. Then, an iterative energy separation algorithm is described which eliminates the need for precise values of formant center frequencies, because the ESA is used to converge to those center values. Next, we describe the system and present experimental results that demonstrate its effectiveness. Application of the system to a vocoder is then discussed. Finally, we conclude and discuss some extensions of our work.

2. BACKGROUND

We define a discrete-time AM-FM signal by

$$x(n) = a(n)\cos(\phi(n)) = a(n)\cos\left(2\pi T \int_0^n f(m)dm\right) \qquad (4)$$

where $|a(n)|$ is the discrete-time amplitude envelope, $f(n)$ is the instantaneous frequency (in Hz), a sampled version of (2), and T is the sampling period. By applying the discrete-time Teager-Kaiser energy operator [4]

$$\Psi_d[x(n)] = x^2(n) - x(n-1)x(n+1) \qquad (5)$$

to the signal $x(n)$ and its backward difference $y(n) = x(n) - x(n-1)$, it has been shown in [1, 2] that

$$\Psi_d[x(n)] \approx a^2(n)\sin^2[2\pi T f(n)] \qquad (6)$$

$$\frac{1}{2\pi T}\arccos\left(1 - \frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4\Psi_d[x(n)]}\right) \approx f(n) \qquad (7)$$

$$\sqrt{\frac{\Psi_d[x(n)]}{1 - \left(\frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4\Psi_d[x(n)]}\right)^2}} \approx |a(n)| \qquad (8)$$

This research work was supported by the National Science Foundation under Grant MIP-91-20624, by a NSF Presidential Young Investigator Award under Grant MIP-86-58150 with matching funds from Xerox, and by MIT Lincoln Laboratory.

0-7803-0946-4/93 $3.00 © 1993 IEEE

Eqns. (7)–(8) are referred to as the discrete-time ESA. At each sample it provides an estimate of the envelope and instaneous frequency using only a 5-sample moving window and at a very small computational complexity. The approximations involved are valid as long as the amplitude envelope and instantaneous frequency do not change too much or too quickly in time compared with the carrier frequency. In implementing the ESA, we also pre-smooth the energy signals $\Psi_d[x(n)]$ and $\Psi_d[y(n)]$ with a 7-point binomial smoothing filter, because this can reduce the approximation errors by about 50%; for details see [5].

3. ITERATIVE ESA

From the results of our early experiments with the ESA, we noticed that when the center frequency of the filter was off by even several hundred Hz, the average value of the instantaneous frequency was often close to the formant peak frequency.

Based on this observation and a suggestion by Kaiser [6], we reasoned that we might be able to use $f(n)$ to iteratively estimate the center frequency of the formant, adjusting the center frequency of the filter on each iteration. Assuming in the AM–FM model (1) for a speech resonance that the frequency modulating signal $q(t)$ has a zero mean within the short-time speech analysis frame, an estimate for the formant center frequency f_c can be the *average* of the instantaneous frequency $f(t)$. Thus, we have implemented the idea of iterative estimation by using the following rule:

$$f_c^{(j+1)} = \frac{1}{N}\sum_{n=0}^{N-1} f^{(j)}(n) \tag{9}$$

That is, the center frequency of the Gabor bandpass filter on the $(j+1)$-th iteration is set equal to the average value of $f(n)$ on the j-th iteration. We start the algorithm by setting $f_c^{(1)}$ to be some initial estimate of the formant, and we consider the algorithm to have converged when the center frequency does not change by more than 5 Hz. This iterative application of the Gabor bandpass filter, the ESA, and the updating of the filter center frequency, while keeping its bandwidth fixed, is henceforth called the ***iterative ESA***.

We have been using the iterative ESA for some time now and, overall, the results are good. For well defined spectral peaks and fairly good initial estimates, the algorithm converges quickly. Fig. 1 shows an example where only one iteration was required for convergence. In this case the initial estimate was off by 140 Hz. A poorer initial estimate or a poorly defined peak requires more iterations.

In Fig. 2 we superimpose the results of the iterative ESA onto the LPC spectrum of a vowel. The LPC spectrum has some peaks that are difficult to distinguish, while the iteration results correspond well with the speech formants. In addition, the iterative ESA has the advantage over LPC that it also finds the modulations, i.e., the signals $|a(n)|$ and $f(n)$. This algorithm seems to be most useful when used in conjunction with a more standard formant finder that provides initial estimates of the formant center frequencies. We will describe such an implementation in Section 4.

We now briefly turn to the issue of how the iterative ESA may be converging. We experimentally observed that the resulting average value of the instantaneous frequency $f(n)$ seemed to be drawn close to peaks or local maxima in the power spectrum. Since the output of $\Psi_d[x(n)]$ is proportional to the energy required to produce $x(n)$, we reasoned that the algorithm could be maximizing the average energy

$$E(j) \quad = \quad \frac{1}{N}\sum_{n=0}^{N-1} \Psi_d\left[x^{(j)}(n)\right] \tag{10}$$

where $x^{(j)}(n)$ is the bandpass filtered speech on the j-th iteration. This quantity was then computed on each iteration while finding the formants of speech segments. We have experimentally found that for the majority of formants $E(j)$ increased as the algorithm converged. We have also found that the average energy E (as a function of the filter center frequency) does peak in the vicinity of formant peaks. This suggests that locally searching for the peaks of E may be useful as a convergence criterion. We are continuing to investigate this issue.

4. AUTOMATED SYSTEM

We now describe an automated system that we have been developing to find the formant center frequencies and modulations of a speech segment. In this system, the iterative ESA, described in the previous section, is employed to determine the center frequencies of the resonances. Bandpass filtering is implemented using a truncated discretized Gabor filter with impulse response

$$g(n) \quad = \quad \begin{cases} \exp(-(\alpha nT)^2)\cdot\cos(2\pi f_c Tn), & |n| \leq N \\ 0, & |n| > N \end{cases} \tag{11}$$

N is chosen so that the envelope of $g(n)$ is nearly zero at $n = N$; we have found that a good choice is N such that $\exp(-\alpha TN)^2 \approx 10^{-6}$. Through extensive experience, we have found that it is reasonable to use fixed bandwidths of $\alpha = 800$ Hz when $f_c < 1000$ Hz, and $\alpha = 1100$ Hz for all other resonances. We discuss the possibility of varying bandwidth values in Section 6.

An important issue for the system was getting good initial estimates of the formant center frequencies. This is to ensure that the iterative ESA does not converge to false formants. We now briefly discuss the two methods that we have implemented: a standard method, LPC, and a new method that we call *morphological peak picking*.

For LPC, the advantages are that is easy to implement and often does a good job of estimating the spectral peaks. However, it sometimes performs poorly, especially for female speakers or children.

The other formant finding method that we have implemented is to perform a morphological closing of the speech spectrum. A *closing* of a signal by a window of W consecutive samples is a nonlinear filter that is a cascade of a *dilation* (local maximum within the moving window) followed by an *erosion* (local moving minimum) [7]. Fig. 3 shows that in a closing, the narrow valleys of the spectrum get filled up, so to speak, and the peaks of the closing correspond to peaks of the spectral envelope. Then it only remains to pick the peaks from the closing and we have our initial formant estimates.

A requirement of this method is that the width, W, of the filter be carefully chosen. It is important that the filter not be too narrow to avoid having too many extraneous peaks, but, at the same time, we must keep it from being too wide to avoid missing a formant that is close to another, stronger formant. Thus, the filter width is essentially a function of the fundamental frequency. We estimate the fundamental frequency by peak picking the spectrum over the first 1000 Hz, and averaging the distances between the resulting harmonic frequencies.

Advantages of morphological filtering are that it is very cheap to implement and can be used rigorously to extract peak or valley features on arbitrary signals. In addition, it is non-parametric, i.e., it does not pre-suppose anything about the speech spectrum, while LPC assumes that the vocal tract transfer function can be modeled by an all-pole model. Finally, it formalizes what we most likely do when we visually identify formants from a speech spectrum using geometrical features.

An example of the output of the automated system is shown in Fig. 5. For this speech segment, the initial estimates of formant center frequency were found using morphological peak picking. Fig. 5(a) shows the speech waveform, while Fig. 5(b) shows the speech spectrum, with the formant center frequencies found by the iterative ESA indicated by vertical dashed lines. Figs. 5(b)-(e) show $|a(n)|$ and $f(n)$ for the third and fourth formants. For this particular example, there are more modulations present in the higher formants than at the lower formants. Median filtering has been applied to the extracted instantaneous frequency signals to suppress the narrow spikes that are due to pitch period effects or isolated numerical instabilities of the ESA.

5. AM–FM MODULATION VOCODER

The iterative ESA is a powerful tool for speech formant tracking and demodulation. In this section we use the iterative algorithm as the analysis part of a novel vocoder.

During real speech analysis and synthesis experiments on voiced utterances using the AM-FM model, we observed significant amounts of amplitude and frequency modulation in speech formants. When the modulations are removed (according to the linear model) speech quality deteriorates considerably. Motivated by the perceptual importance of modulations in speech formants, we introduce in this paper a speech analysis/synthesis system: the *AM-FM modulation vocoder*.

The AM-FM vocoder extracts the *formant bands* from the speech spectrum by filtering the speech waveform around the formant center frequencies. Thus, the center frequencies of the vocoder bands change with time, following the formant variations. Next, each formant band is demodulated into its amplitude envelope and instantaneous frequency components using the discrete-time ESA. The demodulated signals are lowpass filtered, subsampled and coded. At the receiver the information signals are decoded and used for band reconstruction. Finally, the bands are added together to reproduce the speech waveform. The block diagram of the vocoder is shown in Fig. 4.

The automated system presented in the previous section is used for extracting and demodulating the formant bands. The Gabor filter parameters f_c, α are updated at each analysis frame (typically 10 msec).

The AM–FM vocoder combines the advantages of the formant [8] and phase vocoders [9], while avoiding some of their drawbacks. The non-linear second order terms of speech resonances are accounted for, being modeled as amplitude and frequency modulations, while the formant vocoder models only the linear terms. In addition, by choosing the bands to follow the formant variations, the demodulated signals now have the intuitive interpretation of being the envelope and instantaneous frequency of a real speech resonator output, as opposed to the fixed-band phase vocoder.

The AM–FM vocoder, without parameter quantization, has been tested on both isolated voiced phonemes and words with good results. The synthesized speech sounded almost identical to the original utterances. We are currently investigating efficient coding schemes for encoding the amplitude and frequency signals of each resonance.

6. DISCUSSION

In future work, we plan to investigate the following refinements. First, we would like to improve the system by implementing the automatic selection of the filter bandwidths, based on the distance between neighboring formants, because, as mentioned earlier, a badly chosen filter bandwidth may result in exclusion of modulations or inclusion of neighboring formants. We have done some preliminary work on this, using synthetic AM-FM signals, and the results suggest that the optimum choice of bandwidth could be a linear function of the distance between formants, Δf. Our next step is to try and apply this result to speech. Complications may occur, however, when we try to incorporate this with the iterative ESA, because Δf will change during iteration. The solution that we propose is to apply the iterative ESA to all formants in *parallel*. Then we can vary the filter bandwidths on each iteration according to the values of $f_c^{(j)}$.

The other refinement is to possibly reduce discretization effects as follows. Instead of convolving the speech signal s with a discrete-time Gabor bandpass filter g and then applying the discrete-time energy operator, we apply the following combination of the continuous-time energy operator and bandpass filtering, which introduces discretization only at the very last step:

$$\Psi_c[s(t)*g(t)]_{t=nT} = \Big[(s(t)*\dot{g}(t))^2 - (s(t)*g(t))(s(t)*\ddot{g}(t))\Big]_{t=nT} \quad (12)$$

where $\dot{g}(t)$ and $\ddot{g}(t)$, the derivatives of the Gabor bandpass filter, are functions with simple known formulas. In this way, we avoid the approximation of the signal derivatives with first differences which maps Ψ_c to Ψ_d [1], and this may improve the results of applying the energy operator and the ESA to sampled speech signals.

Acknowledgment: We wish to thank Jim Kaiser of Rutgers University and Tom Quatieri of MIT Lincoln Labs for many insightful discussions.

REFERENCES

[1] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," *Proc. IEEE ICASSP-91*, Toronto, Canada, pp. 421–424, May 1991.

[2] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators," *Proc. IEEE ICASSP-92*, San Francisco, Calif., pp. II-1-4, Mar. 1992.

[3] J. F. Kaiser, "On Teager's energy algorithm and its generalization to continuous signals," *Proc. 4th IEEE Digital Signal Processing Workshop*, Mohonk (New Paltz), NY, Sept. 1990.

[4] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. IEEE ICASSP-90*, Albuquerque, NM, pp. 381–384, April 1990.

[5] A. Potamianos and P. Maragos, "A comparison of the energy operator and Hilbert transform approaches to signal and speech demodulation," Tech. Report 92-8, Harvard Robotics Lab, July 1992.

[6] J. F. Kaiser, personal communication, 1991.

[7] P. Maragos and R. W. Schafer, "Morphological filters—part I: their set-theoretic analysis and relations to linear shift-invariant filters," *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, ASSP-35, pp. 1153–1169, Aug. 1987.

[8] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971–995 Mar. 1980.

[9] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Am.*, vol. 68, pp. 412–419, Aug. 1980.

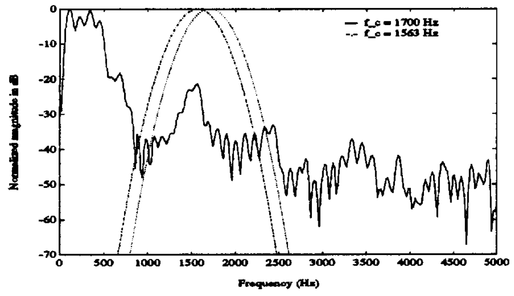


Figure 1. *The iterative ESA was started at $f_c^{(1)} = 1700$ Hz for this speech segment. Only one iteration was necessary for it to converge to $F2 = 1563$ Hz, a difference of about 140 Hz.*

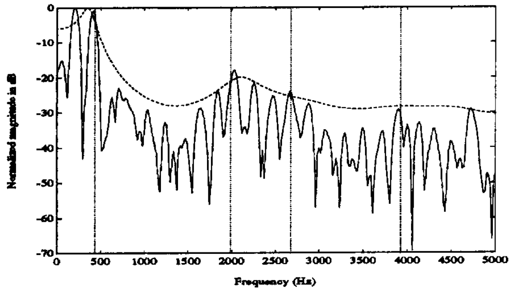


Figure 2. *The solid and dashed lines represent the speech and LPC spectra, respectively, of 20 msec of the vowel /u/, spoken by a female. The vertical dotted lines indicate the formant center frequencies found by the iterative ESA.*

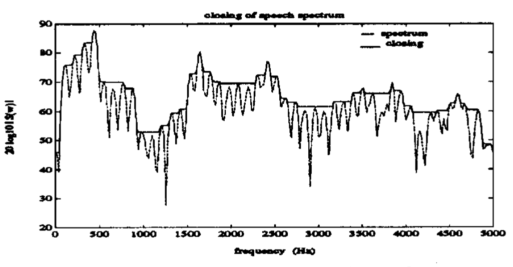


Figure 3. *Morphological closing of speech spectrum.*

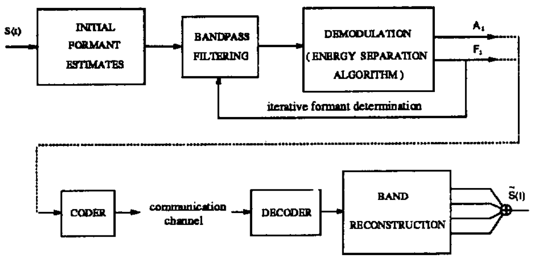
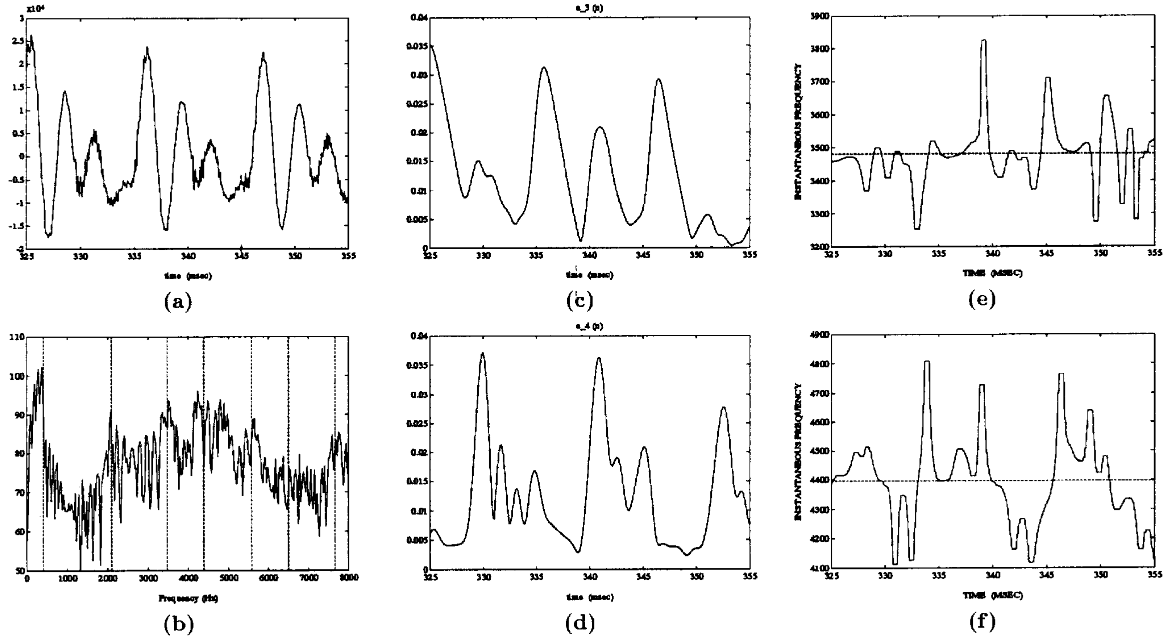

S(t)
INITIAL FORMANT ESTIMATES
BANDPASS FILTERING
DEMODULATION (ENERGY SEPARATION ALGORITHM)
A_i
F_i
iterative formant determination
CODER
communication channel
DECODER
BAND RECONSTRUCTION
Ŝ(t)

Figure 4. *Block diagram of the AM–FM vocoder.*



(a) (c) (e)

(b) (d) (f)

Figure 5. *Results of the automated system. (a) The vowel /i/, from a male speaker. (b) The spectrum. Dotted vertical lines indicate formant center frequencies following iteration. (c)–(d) Amplitude envelopes $|a(n)|$ for the 3rd and 4th formants, respectively. (e)–(f) Instantaneous frequencies $f(n)$ for the 3rd and 4th formants, respectively, after 13 point median filtering.*