A ROBUST METHOD FOR THE ESTIMATION OF FORMANT FREQUENCY MODULATION IN SPEECH SIGNALS

Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Kanpur 208016 INDIA
(preeti@iitk.ernet.in)

ABSTRACT

Recently the presence of amplitude and frequency modulations in individual formants of the speech signal was demonstrated using the Discrete-time Energy Separation Algorithm (DESA). Formant modulation estimates are valuable to the understanding of speech production, and have found several applications. While the DESA has been successfully applied in tracking the amplitude envelope of formants, the DESA frequency estimator has been relatively unexploited for tracking formant frequency modulation (FM) due to its lack of robustness to conditions commonly occurring in practice. Here we consider an alternate method of frequency estimation based on the Wigner Distribution (WD) and investigate its application to the estimation of formant FM in speech signals. It is shown using simulated and speech signals that the WD estimator can track formant FM with high time resolution under a wider range of conditions and is more robust to additive noise present in the signal than the DESA estimator. Finally the computational complexities of the two methods are compared.

1. Introduction

Recently the presence of amplitude and frequency modulations in individual formants of the speech signal was demonstrated using the Discrete-time Energy Separation Algorithm (DESA) [1]. Time variations of instantaneous frequency and amplitude were observed within individual pitch periods of bandpass filtered speech resonances, supporting theories of nonlinear and time-varying phenomena during speech production. Based on this, the following exponentially-damped AM-FM signal has been proposed as a model for a speech resonance [1],

$$\begin{aligned} x(n) &= a(n)cos[\phi(n)] \\ &= r^n A(n)cos\left[2\pi f_c n + 2\pi f_m \int_0^n q(k)dk + \phi(0)\right] \end{aligned} \quad (1)$$

with a time-varying amplitude $a(n)$ and an instantaneous frequency given by $f_i(n) = f_c + f_m q(n)$. The speech signal is modeled as the sum of several such resonances, one corresponding to each formant frequency. With a bandpass filter, centered at the formant center frequency, applied to isolate the resonance, the DESA has been used to estimate separately the amplitude and frequency signals, $a(n)$ and $f_i(n)$, from sampled speech. Apart from the value of such investigations to the understanding of speech production, formant modulation information has found applications in speech coding, speech classification and speaker identification [2,3].

For a discrete-time AM-FM signal of the form (1), the DESA provides a computationally simple, high resolution estimate of the envelope and instantaneous frequency under certain assumptions. While the amplitude envelope is accurately tracked under most conditions occurring in practice, the DESA frequency estimator breaks down when the assumptions necessary for its validity do not hold, such as when the carrier frequency is low relative to the maximum frequency deviation, and in the presence of additive noise. This has limited the use of formant frequency modulation estimates in practical applications [2]. It is of interest, therefore, to investigate the application of alternate demodulation approaches to estimate instantaneous frequency (i.f.).

One obvious method to estimate the i.f. of phase-modulated signals is by the phase-derivative of the corresponding analytic signal obtained using the Hilbert Transform (HT) [1]. While this method provides accurate estimates of i.f. in a wider range of modulation conditions than the DESA frequency estimator, it is not robust to additive noise present in the signal.

The Wigner Distribution (WD), a joint time-frequency signal representation, offers a representation concentrated about the instantaneous frequency for phase-modulated signals. Further, the frequency location of the peak of the WD at each time instant provides a nearly optimal estimate of the i.f. for a large

0-7803-3192-3/96 $5.00©1996 IEEE

class of phase-modulated signals in additive white, Gaussian noise (w.g.n.) [4]. The WD has been applied to the tracking of i.f. of time-varying signals in several applications including the estimation of formant tracks during transient sounds in speech. Here we discuss the application of the WD to estimate formant frequency variation at far smaller time scales i.e., on the order of a sampling period. We compare its performance with the DESA and HT frequency estimators using simulated as well as speech samples.

The DESA frequency estimator is given by the DESA-1 of [1],

$$\begin{aligned} f_i(n) &= \frac{1}{2\pi} arccos \left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]} \right) \\ y(n) &= x(n) - x(n-1) \\ \psi[x(n)] &= x^2(n) - x(n-1)x(n+1) \qquad (2) \end{aligned}$$

2. Implementation of the WD frequency estimator

For a complex signal s(n), the discrete-time WD is given by

$$W_s(n,f) = \sum_{k=-N/2}^{N/2} s(n+k)s^*(n-k)\exp(-j4\pi kf) \quad (3)$$

The above is computed via a highly zero-padded DFT of the windowed inner product sequence at each time sample n. The resulting function of f is searched for its maximum value. The i.f. estimate is the corresponding value of f. For real signals of the form (1) the analytic form is first computed using the Hilbert transform and then (3) is implemented. Typically a Gaussian window is applied to the data before computing the WD to optimise the obtainable time- and frequency- resolutions.

For complex signals with constant amplitude and quadratic phase function (linear FM) in the presence of additive w.g.n., the frequency location of the peak of the WD, as computed in (3), gives the optimal estimate of the i.f. at the time instant n [4]. The estimate is unbiased with the variance decreasing with increasing window length N. In the case of nonlinear FM signals, the WD frequency estimate is generally biased with the amount of bias depending on the deviation from linearity (or more accurately, skew-symmetry) of the i.f. curve within the data window. Thus the choice of window length (and shape) reflects the trade-off between bias and variance of the i.f. estimate.

3. Performance on Simulated Signals

The DESA and WD estimators have been applied to the estimation of instantaneous frequency for a class of AM-FM/Cosine signals which serve as a realistic model for the behaviour of a single speech resonance within a pitch period [1]. The set of signals used in the simulation are given by:
$[1 + 0.5cos(\pi n/100)]cos(\pi n/5 + 20k sin(\pi n/100))$, with k varying in (0.5,5) in steps of 0.5, and n : 0..999. The DESA algorithm is implemented as in (2) followed by 5-point median filtering [1]. "WD-N" is the WD frequency estimate computed from the analytic signal obtained via frequency domain filtering; N is the length in samples of the Gaussian window applied to the data. The size of the DFT in the WD computation is 512.

Table 1. Percentage frequency estimation errors for AM-FM/Cosine signals in additive w.g.n.

SNR	DESA		WD-16		WD-32	
(dB)	abs	rms	abs	rms	abs	rms
Infinity	0.3	0.4	0.05	0.09	0.16	0.23
20	10.5	18.0	1.2	1.6	0.47	0.64
10	31.6	48.0	5.5	14.0	1.7	5.8

From the results of the numerical simulation in Table 1 we see that for the given window lengths, the frequency estimate based on the WD exhibits less bias than the DESA frequency estimate (as depicted by the values at SNR=Infinity). In the presence of additive w.g.n., the WD frequency estimate is significantly more accurate than the DESA frequency estimate. Increasing the window length in the WD estimator, while increasing the bias, improves its robustness to noise. It was found that frequency estimation based on the HT exhibited a bias and variance comparable to that of the DESA in Table 1.

4. Application to Speech Analysis

The WD frequency estimator as described in Section 2, has been applied to the tracking of instantaneous frequency of bandpass filtered speech formants and the resulting estimates compared with those obtained by the DESA estimator and the HT-based estimator. Since the WD is based on the analytic signal, its performance on noise-free signals is expected to be similar to that of the HT-based estimator. We have made the following general observations on the comparative behaviour of the three frequency estimators. Some of these are demonstrated by the example of Figure 1 for a segment of the vowel "ae" (male speaker) spanning about 3 pitch periods and sampled at 16 KHz. The frequency estimation algorithms were applied to the bandpass filtered formants F1 (700 Hz) and F3 (2900 Hz). Gaussian filters centered at the formant frequencies, and bandwidths of 900 Hz and 1100 Hz respectively, were used.

4.1. Dependence on formant frequency:

For middle and high frequency formants (frequency>1 KHz) the WD frequency estimate computed using a short Gaussian-shaped window (length = 16 samples at 16 KHz sampling frequency) matches the corresponding DESA frequency estimates closely, i.e.,the WD estimate tracks the formant frequency modulation with high time resolution between pitch impulses while exhibiting sharp spikes due to phase discontinuities at the pitch period boundaries (see Fig.1 c,d). The amplitude of the spikes can be reduced by increasing the WD window length but this also results in smoothing of the formant FM due to the increased time-averaging as seen in Fig 1e. However this additional flexibility provided by the availability of the window length parameter in the WD estimator sometimes improves the overall representation of the FM. For low frequency formants, the DESA estimate is significantly degraded since the assumption that the carrier frequency is low relative to the frequency deviation is no longer valid. The WD and HT estimates shows formant FM clearly within each pitch period (note the 2 lobes per pitch interval in Fig 1 g,h).

4.2. Effect of additive noise:

From the simulation results of Table 1, it is evident that the WD frequency estimator is far more robust than the DESA (or HT) estimator in the presence of additive, white noise. This characteristic is useful in regions of low amplitude of the input signal, when noise due to quantisation in the digital computation can be significant. It is found that while the DESA and HT estimators often break down in low amplitude regions of the input speech signal, the WD estimate tracks the frequency accurately.

Since frequency estimation is done after applying a bandpass filter to the speech signal to suppress neighbouring formants, any additive noise present in the speech signal is considerably attenuated. Therefore the effect of additive noise in the speech signal on the estimation of formant FM is dependent on the bandwidth of the bandpass filter, and for typical filter bandwidths, can be observed only at very low overall speech SNRs. In Fig. 1 (i,j) the WD and HT estimators are applied to the frequency estimation of formant F1 of the speech signal in additive w.g.n. at the very low overall SNR of 3 dB (before bandpass filtering). With the given window length (48 samples) we see that the WD estimate is better able to preserve the formant FM in the presence of noise.

4.3. Computational complexity:

The DESA frequency estimator in (2) is computed using only 5 data samples at any time instant and is computationally very simple. The WD estimator on the other hand, involves computation of an inner product followed by a large (but sparse) DFT, and hence is far more complex than the DESA estimator. However in the application to speech formant FM estimation, the computation necessitated by the bandpass filter required to isolate the formant adds considerably to the overall complexity and makes the two algorithms a little more comparable. In the case of the WD estimator, the analytic signal can be generated by applying a complex bandpass filter to the speech signal, which requires only twice as much computation as the real bandpass filter required when using the DESA estimator. Further, in implementing (3) to find the frequency maximum of the WD, it is found necessary to search only the immediate neighbourhood (+/-10 frequency samples) of the previous frequency peak in order to find the current frequency maximum, limiting the computation to about 20 DFT samples. The overall complexity of the WD-based algorithm in the speech formant FM estimation application is typically 5-10 times higher than that of the DESA frequency estimator.

References

[1] Maragos,Kaiser and Quatieri, "Energy Separation in Signal Modulations with Application to Speech analysis," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3024-3051, 1993.

[2] C.R.Jankowski Jr.,T.F.Quatieri and D.A.Reynolds, "Formant AM-FM for Speaker Identification," *Proc. of the IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*, 1994.

[3] Potamianos and Maragos, "Speech Formant Frequency and Bandwidth Tracking using Multiband Energy Demodulation," *Proc. ICASSP*, 1995.

[4] P.Rao and F.J.Taylor, "Estimating Instantaneous Frequency using the Discrete Wigner Distribution," *Electronics Letters*, vol. 26,no. 4, 1990.

Acknowledgement

Part of this work was carried out at the C.S.C. Group, T.I.F.R., Bombay, under a Visiting Fellowship from the Indian National Science Academy. Valuable discussions with Prof.P.V.S.Rao and Dr.K.Samudravijaya are gratefully acknowledged.

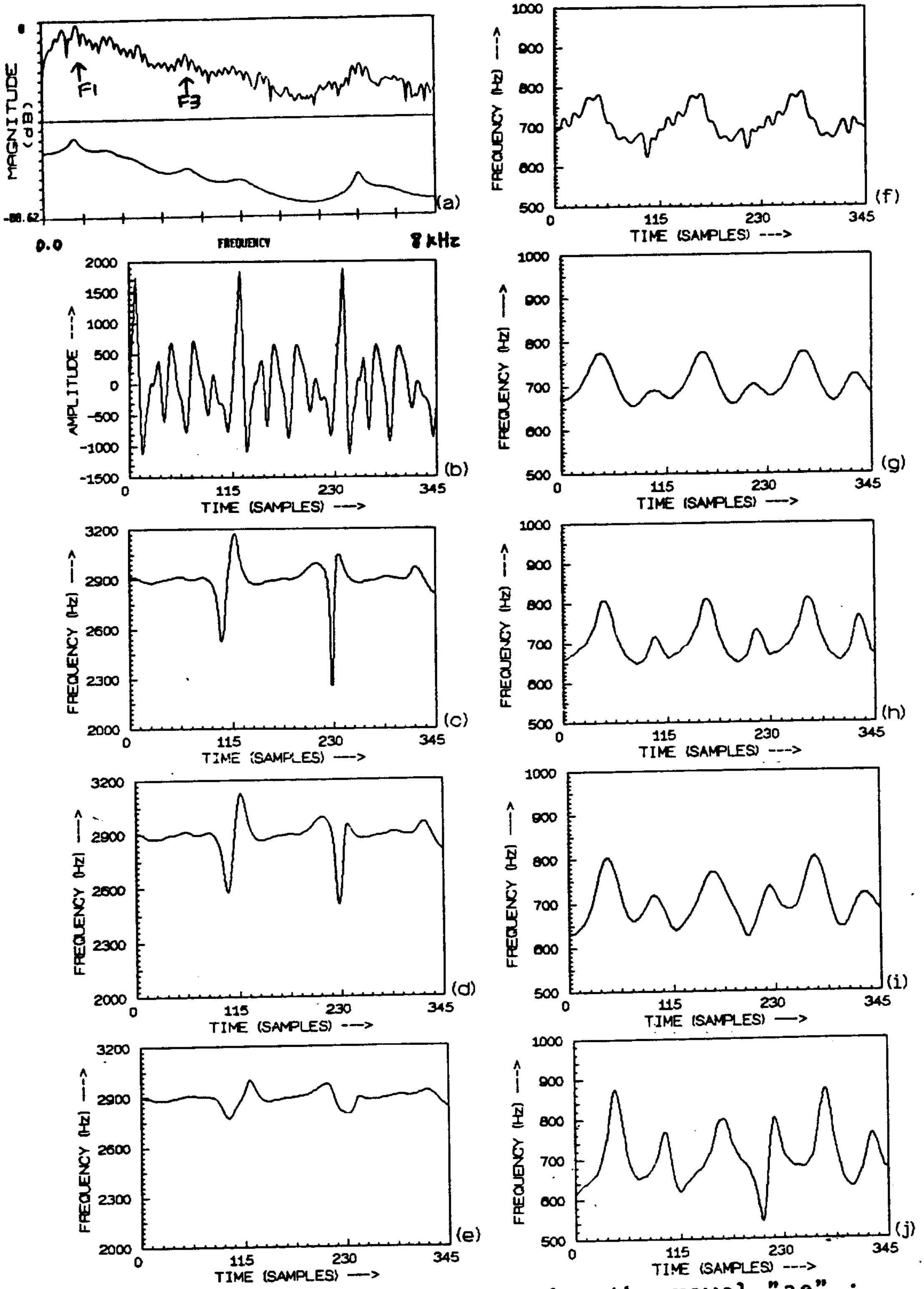

Fig.1(a) Speech and LPC spectra for the vowel "ae" ;
(b) Time waveform of "ae" (male voice,samp.freq.=16KHz);
Frequency estimates for F3:(c) DESA;(d) WD-16;(e) WD-48;
Frequency estimates for F1:(f) DESA;(g) WD-48;(h) HT
Frequency estimates for F1 (SNR=3dB): (i) WD-48;(j) HT