ELSEVIER

Signal Processing 80 (2000) 1655–1667

SIGNAL PROCESSING

www.elsevier.nl/locate/sigpro

Speech formant frequency estimation: evaluating a nonstationary analysis method

Preeti Rao[a,b,*], A. Das Barman[a]

[a]*Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208016, India*
[b]*Advanced Centre for Research in Electronics, Indian Institute of Technology, Bombay, Powai, Mumbai 400076, India*

Received 22 April 1999; received in revised form 16 March 2000

Abstract

The objective of this paper is to critically evaluate the performance of a nonstationary analysis method in tracking speech formant frequencies as they change with time due to the natural variations in the vocal-tract system during speech production. The method of instantaneous frequency estimation is applied to the tracking of speech formant frequencies to observe the time variations in the vocal-tract system characteristics within a pitch period. An implementation of an instantaneous frequency estimator based on the source–filter model of speech production is described for voiced speech formants. Based on experimental results from simulated as well as natural speech data, it is shown that the accuracy of the frequency estimates is heavily dependent on the nature of the glottal excitation waveform, the fundamental frequency and the frequency spacing of the formants in the speech signal. The choice of various analysis parameters on the accuracy of the estimates is discussed. It is shown that only when the formants are well separated and there are distinct regions of the glottal cycle in which the source excitation can be considered to be negligible, does the instantaneous frequency estimate accurately represent the actual formant frequency. Experimental results on natural speech vowels which show differences in formant frequencies in the different phases of the glottal cycle are presented.

Zusammenfassung

Das Ziel dieses Artikels ist es, die Leistungsfähigkeit einer nichtstationären Analysemethode zum Verfolgen der zeitlichen Änderungen der Formantfrequenzen, wie sie durch die natürlichen Variationen des Vokaltraktes während der Spracherzeugung entstehen, zu evaluieren. Um Schwankungen in der Charakteristik des Sprachtrakts innerhalb einer Pitch-Periode zu beobachten wird die Methode der Schätzung der Momentanfrequenz zum Verfolgen der Formantfrequenzen angewandt. Eine Implementierung der Momentanfrequenzschätzung für stimmhafte Formanten, basierend auf dem Quelle–Filter-Modell der Spracherzeugung, wird beschrieben. Basierend auf experimentellen Ergebnissen sowohl mit simulierten als auch mit natürlichen Sprachsignalen wird gezeigt, dass die Genauigkeit der Frequenzschätzungen stark von der Art des Glottis-Signals, der Grundfrequenz und dem Frequenzabstand der Formanten im Sprachsignal abhängt. Der Einfluss der Wahl verschiedener Analyseparameter auf die Genauigkeit der Schätzung wird diskutiert. Es wird gezeigt, dass nur bei gut separierten Formanten und bei ausgeprägten Bereichen des Glottis-Zyklus, in denen das Anregungssignal als vernachlässigbar betrachtet werden kann, die Schätzung der Momentanfrequenz die tatsächlichen Formantenfrequenzen richtig wiedergibt. Experimentelle Ergebnisse mit natürlich erzeugten Vokalen, die Unterschiede in den Formantfrequenzen während der unterschiedlichen Phasen des Glottis-Zyklus aufweisen, werden vorgestellt.

* Corresponding author. Tel.: + 91-22-576-7695; fax: + 91-22-572-3806.
E-mail address: prao@acre.iitb.ernet.in (P. Rao).

0165-1684/00/$ - see front matter
PII: S0165-1684(00)00099-2

Résumé

L'objectif de cet article est l'évaluer de façon critique les performances d'une méthode d'analyse non stationnaire dans la poursuite de fréquences de formants de parole alors qu'elles changent au cours du temps à cause des variations naturelles du système du conduit vocal pendant la production de parole. La méthode d'estimation de la fréquence instantanée est appliquée à la poursuite de fréquences de formants de parole pour observer les variations temporelles dans les caractéristiques du système de conduit vocal à l'intérieur d'une période de ton. Nous décrivons une implémentation d'un estimateur de fréquence instantanée basée sur un modèle se filtre de source de la production de parole, pour des formants de parole voisée. Sur la base de résultats d'expériences à la fois pour des données de parole simulée et naturelle, nous montrons que la précision des estimations de fréquences est fortement dépendant de la nature de l'onde d'excitation glottale, de la fréquence fondamentale et de l'espacement fréquentiel des formants dans le signal de parole. Le choix de différents paramètres d'analyse sur la précision des estimées est discuté. Nous montrons qu'uniquement lorsque les formants sont bien séparés et qu'il existe des régions distinctes du cycle glottal dans lesquelles l'excitation de source peut être considérée comme négligeable, l'estimation de la fréquence instantanée représente de façon précise la fréquence réelle des formants. Nous présentons des résultats expérimentaux sur des voyelles de parole naturelle qui montrent des différences en fréquences de formants dans différentes phases du cycle glottal.

Keywords: Speech analysis; Formant tracking; Instantaneous frequency

1. Introduction

Estimation of vocal-tract characteristics from the speech signal has been an important research topic due to its usefulness in understanding and modelling the speech production mechanism. The properties of the vocal tract vary in time due to the movement of the articulators, and due to vocal fold oscillations in each pitch period [1,11]. In the analysis of speech signals, formant parameters are commonly used to characterise the vocal tract. Due to the movement of articulators during speech production, the formant parameters vary with time. These variations are usually slow except in the case of certain speech sounds of a highly dynamic nature. The relatively more rapid variations in vocal-tract characteristics due to vocal fold oscillations can be understood by considering the widely used source–filter model for voiced speech. A source that generates a sequence of glottal pulses is used to model the air flow. The radiation effects of the lips are incorporated by replacing the excitation signal to the vocal tract by a sequence of time-differentiated glottal pulses. Fig. 1 shows the Liljencrants–Fant (LF) model differentiated glottal pulses based on a typical set of parameters [1]. Since glottal closure is usually abrupt, there is a virtual absence of source excitation (time derivative of the glottal pulse) in the closed phase of the glottal cycle. This leads to free resonances in the speech signal, corresponding to the formant frequencies. In the open phase the vocal folds and the tract are acoustically coupled, and the source excitation is typically non-zero. The coupling of the source and tract during the open phase leads to changes in the bandwidths

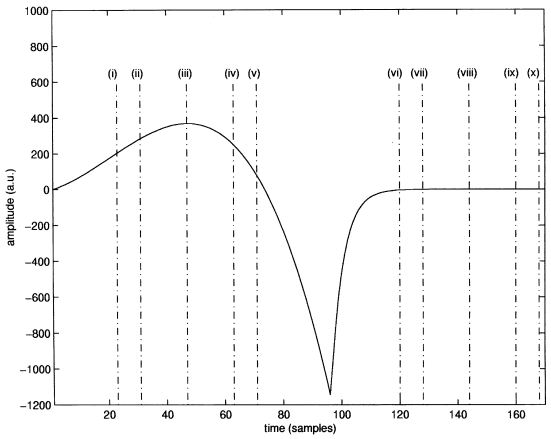

Fig. 1. LF model glottal pulse time-derivative waveform (pitch period = 10.6 ms). The dotted lines indicate the time instants referred to in the rest of the paper.

and frequencies of the resonances of the vocal tract [1].

Traditionally, quasistationary methods using block-based processing have been used to track time variations in formant parameters. The performance of two quasistationary methods, namely the spectrogram and linear prediction, has been comprehensively investigated and reported in [7,11]. The accuracy of the analysis in the estimation of rapidly varying formants has been quantitatively evaluated with respect to the window length and position. It has been observed that the most accurate analysis using a quasistationary method is made when the effective length of a window is smaller than a pitch period and windows are positioned pitch synchronously. However, observing variations within a pitch period due to the different phases of the glottal cycle necessitates the use of analysis frames matched to the duration of each phase which may be much less than a pitch period. High-resolution spectrum estimation techniques such as covariance-based linear predictive analysis have been applied to short segments of data lying within each phase using pitch-synchronous positioning of windows [11].

In this paper we investigate a nonstationary analysis approach to observe formant frequency changes in voiced speech at time resolutions which are much less than a pitch period. These variations may be due to source–tract interaction and possibly also due to the dynamics of the underlying sound. Nonstationary analysis techniques have been applied to speech representation in the past. In particular, time–frequency distributions have been widely used to present graphically the broad time evolution of time-varying signal characteristics. While such representations provide a rich level of detail, they can often be difficult to interpret. An alternative approach to nonstationary signal analyses is to use a particular property of the time–frequency distribution to obtain the evolution with time or frequency of a useful signal or system parameter. Instantaneous frequency is one such parameter which can be derived from a time–frequency analysis of the signal. Since we are interested in tracking the time variation of formant frequencies of the speech signals, it is natural to consider the application of instantaneous frequency (IF) estimation. Such IF estimates obtained for each time-varying resonance of the vocal tract can possibly be used to construct a spectrogram-like display of formant tracks at a high time resolution.

Instantaneous frequency estimation has been previously applied to bandpass filtered speech signals to track formant frequency variation with high time resolution [4]. The interpretation of the estimates however has been based on the assumption that the signal being analysed consists of free resonances of a time-varying vocal-tract system (that is, the source excitation is modelled by a train of impulses). The instantaneous frequency has been shown to exhibit significant time variation within a single pitch period. However, the assumption that source excitation is absent holds at best only for a fraction of the glottal cycle duration in voiced speech. The presence of source excitation during the glottal cycle has a significant effect on the accuracy of the estimates. In this work, the performance of instantaneous frequency estimation is examined with specific reference to the source–filter model of speech production where the source excitation is modelled by differentiated glottal pulses given by the LF model. An instantaneous frequency estimator based on the Wigner–Ville distribution (WD) is described. This method is representative of this class of nonstationary analysis techniques and more importantly, provides a general framework in which the influence of various analysis parameters can be evaluated. Since a general analytical procedure is not expected to provide the necessary insight on the behaviour of IF estimation on real speech signals, we follow an empirical approach based on experimental investigations using simulated and natural speech data. In the remaining sections of the paper, instantaneous frequency estimation is introduced and the WD-based method is described. The application of IF estimation to speech formant tracking is discussed. The accuracy of the IF estimates is investigated using synthetic signals and the influence of signal characteristics such as the glottal source waveform, fundamental frequency and the frequency spacing of formants on the accuracy of the estimate is quantified with respect to various analysis parameters. Finally, some results of IF estimation applied to natural speech vowels are presented.

2. Instantaneous frequency estimation and the Wigner–Ville distribution

For narrowband phase-modulated signals, the instantaneous frequency (IF) is a useful quantity given by the derivative of the phase of the corresponding analytic signal [8]. For

$$s(t) = a(t)\cos \Phi(t), \tag{1}$$

we have instantaneous frequency given by

$$\omega(t) = \frac{\mathrm{d}\Phi(t)}{\mathrm{d}t}. \tag{2}$$

IF estimators may be computed directly using this definition by first obtaining the analytic signal via the Hilbert transform and then differencing successive phase samples. Variants of this approach have also been proposed such as the T–K algorithm [4]. In the presence of additive noise however, differencing successive samples of a noisy phase sequence exhibits a high statistical variance even at high-input SNR. Hence, these estimators are not robust in the presence of noise and as a consequence also break down in low-amplitude regions of the speech signal [5]. Additionally, the T–K algorithm is not reliable at low formant frequencies due to the breakdown of an assumption necessary for its validity at these frequencies [4]. It has been of interest therefore to investigate alternative methods of IF estimation for application to speech. Recently, an IF estimator based on the phase derivative of the flat-envelope residual obtained by fitting a time-domain pole-zero model to the signal, was proposed for speech applications [3]. This estimate has been shown to have the advantage of providing a positive-valued IF (PIF) estimate for an important class of signals. While the PIF estimate is expected to match the accuracy of other IF estimators in the case of purely phase-modulated signals, its interpretation has not been investigated for the case of signals produced by the source–filter type of model. Also, since the PIF estimate is based on direct differencing of successive phase samples, it is expected that its sensitivity to noise will be high.

The Wigner–Ville distribution (WD), a joint time–frequency signal representation offers a representation concentrated about the instantaneous frequency for phase-modulated signals. For a complex signal $s(n)$, the discrete-time WD is given by

$$W_s(n,f) = \sum_{k=-M/2}^{k=M/2} s(n+k)s^*(n-k)\exp(-\mathrm{j}\pi kf). \tag{3}$$

The above equation can be realised via a highly zero-padded DFT of the windowed inner product sequence at each time sample n. Due to the finite data window in Eq. (3), the resulting function is a frequency-smoothed version of the true signal WD. When applied to realistic multicomponent signals such as speech, the WD displays interference terms. Smoothing along the time and frequency axes is used to effectively suppress these components at the cost of a corresponding loss in resolution [10].

For the case of monocomponent, complex signals with constant amplitude and quadratic phase functions (linear frequency modulation), the frequency location of the peak of the discrete WD (DWD) function at a given time instant n is equal to the instantaneous frequency. In the presence of additive white Gaussian noise, the peak of the DWD provides the optimal estimate of instantaneous frequency at the time instant n for such signals [6]. In the case of nonlinear frequency-modulated (FM) signals, the WD peak frequency estimate is generally biased with the amount of bias depending on the deviation from linearity of the instantaneous frequency within the data window. For a slowly varying FM signal the window length can be chosen small enough so that the IF trajectory is close to linear within the data window. Reducing the data-window length however leads to an increase of the main lobe width corresponding to the DWD peak and hence leads to an increase in the variance of the estimate in the presence of noise. Typically, a Gaussian data window is applied to optimise the obtainable time and frequency resolutions, or alternatively to optimise the trade off between the bias and the variance of the IF estimate. A Gaussian data window tapers off to zero at the extreme samples and has an effective length of half its total duration.

Hence, we note that when the data-window length is small enough so that the signal is well

approximated by a linear FM law within the data window, the WD method gives equivalent results for noise-free signals as the direct computation of the phase derivative while significantly improving the performance of IF estimation in the presence of noise.

3. Application to speech formant frequency tracking

While we have discussed the performance of the DWD peak in the estimation of IF for monocomponent, phase-modulated signals, the speech signal is a far more complicated, multicomponent signal. Based on the source–filter model of speech production, the speech signal can be viewed as the sum of filtered components, each generated by passing the source excitation signal through a single resonance of the vocal-tract filter. For voiced speech the source excitation is well modelled by the time derivative of the glottal-pulse train at the specified fundamental frequency. The vocal-tract system resonances are parameterised by their centre frequencies and bandwidths. In order to apply IF estimation to observe the time variations of vocal-tract resonance frequencies, the speech signal must be decomposed into monocomponent signals, each of which represents the response of a single resonance of the vocal tract. Further the effect of source excitation should be minimised as far as possible. Only then can the phase derivative of the corresponding analytic signal be interpreted as the "instantaneous frequency" of the vocal-tract resonance. In this section we describe the necessary processing of the speech signal in order to apply IF estimation to tracking the formant frequencies in a voiced sound.

Referring once again to Fig. 1 of a typical source excitation waveform given by the LF model glottal-pulse time derivative, we see that while the source excitation is significant at the instant of glottal closure, there is a nonzero excitation present throughout the open phase. The source excitation waveform which dominates the open-phase region is a predominantly lowpass signal and can be suppressed to some extent by pre-emphasis of the speech signal given by $s(n) - s(n-1)$ [11]. This operation serves to localise the source excitation to regions near the instants of glottal closure and opening. To estimate the instantaneous frequency of the individual formants, the pre-emphasised signal must be separated into single-component signals, one for each formant frequency. Bandpass filtering is used for the purpose with the centre frequency of the bandpass filter corresponding to the approximate formant centre frequency. The formant centre frequencies may be identified from an inspection of the average magnitude spectrum of the speech signal computed using a long (say 30 ms) data window. Alternatively a linear prediction spectrum may be used, making it easier to pick the peaks corresponding to formants. The bandwidth of the bandpass filter should be chosen such that any short-time frequency variation of the formant is captured while the neighbouring formants are sufficiently attenuated. While this is not always possible, there are a large number of voiced sounds in which the formants are sufficiently separated to permit this. As in [4] a Gabor filter is used to provide a Gaussian-shaped frequency response which has an optimal joint time and frequency width. Further, using a complex-valued discrete filter allows the analytic signal to be directly generated for the IF estimation algorithm [8].

A complex discrete Gabor bandpass filter can be realised by the following FIR filter:

$$h(n) = \exp(-(\alpha T n)^2)\cos(2\pi f_c n), \quad -N \leqslant n \leqslant N, \tag{4}$$

where T is the sampling period. The length of the filter is $2N+1$ samples and is chosen such that the Gaussian envelope samples are close to zero at the edges. The filter 3 dB bandwidth is given by $\alpha/\sqrt{2\pi}$. The filtering leads to a time-domain smoothing of the signal and is expected to result in the temporal smoothing of the IF estimate. Hence, it is of interest to minimise the length of the filter impulse response, or equivalently, to maximise its bandwidth. In the analysis of speech signals, the frequency proximity of the formants plays a significant role in the choice of filter bandwidth. Typically, a filter bandwidth limited to 400–500 Hz is necessary to suppress the neighbouring formants sufficiently. The complex output of the bandpass filter is used in the DWD computation of Eq. (3) and the frequency

peak of the resulting function is taken to be the estimate of instantaneous frequency. The finite data window in the DWD computation leads to a frequency smoothing of the DWD spectrum. The DWD peak estimate is biased to the extent that the signal within the data window deviates from a linear frequency-modulated signal of constant amplitude. The pre-emphasised and bandpass-filtered speech signal qualifies as a narrowband phase-modulated signal only whenever the source excitation is negligible. The time instants corresponding to the presence of significant excitation such as the opening or closing of the glottis are marked by abrupt phase changes in the signal and hence cannot be expected to yield reliable estimates of the IF of the resonance.

The accuracy of the estimates is influenced by the bandpass filter bandwidth and by the choice of DWD data-window length. The data window length, $M+1$, in Eq. (3) is a parameter of the analysis and the various considerations that govern its choice will be discussed in the succeeding sections.

4. Simulation studies

In this section we present examples which illustrate the behaviour of instantaneous frequency estimation applied to time-varying single- and multi-formant signals. The aim of this simulation study is to gain an understanding of the limitations of the IF estimation methods when applied to realistic speech data and the influence of analysis parameters such as the filter bandwidth and data-window duration on the accuracy of the estimates.

The signals are generated by exciting a time-varying IIR filter with a glottal-pulse time-derivative source excitation given by the LF model. Fig. 1 shows a typical waveform computed from the Rosenberg approximation to the LF model of a glottal-pulse time derivative with a distinct closed phase [9]. A sampling frequency of 16 kHz is used throughout. To study the accuracy of the IF estimation method, the IF estimate is recorded at selected time instants in the glottal cycle period. These time instants are shown in Fig. 1. In the closed phase, we consider the location at the centre (marked (viii)) and two instants on each side of this location spaced at 1 and 1.5 ms from it. In the open phase, the centre location is chosen as the point of minimal excitation in the pre-emphasised glottal pulse time-derivative signal (marked (iii)). The remaining locations are chosen on either side of this sample spaced 1 and 1.5 ms from it.

4.1. Single-formant signals

4.1.1. Method of signal generation

A set of single-formant signals was generated by exciting a second-order IIR filter with the LF model glottal-pulse time derivative. The fundamental frequency of the excitation was set at 94.1 Hz corresponding to a pitch period of 170 samples. The coefficients of the IIR filter were modified at every time sample so that various conditions on the time variation of the resonance frequency could be simulated. The following different signals were simulated.

SS1. Fixed resonance frequency $f = 2000$ Hz.
SS2. Resonance frequency varying linearly at 8 Hz/ms with a centre frequency of 2000 Hz.
SS3. Resonance frequency varying linearly at 32 Hz/ms with a centre frequency of 2000 Hz.

Fig. 2 shows the glottal-pulse time-derivative excitation and the single-formant synthesised signal

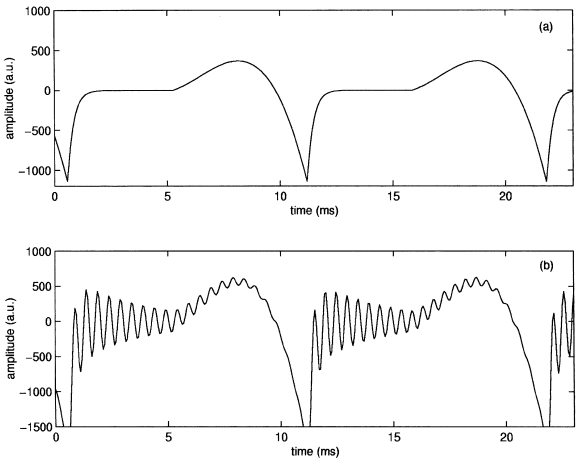

Fig. 2. (a) Source excitation at fundamental frequency = 94.1 Hz; (b) single-formant synthetic signal with fixed resonance frequency = 2000 Hz.

SS1 over a duration of about 2 periods. To study the performance of the IF estimator in noise, zero-mean white Gaussian noise sequences were added to the signal SS1 at various signal-to-noise ratios.

4.1.2. Method of analysis

The synthetic single-formant signal is pre-emphasised and bandpass filtered as discussed in Section 2. The complex bandpass filter has a centre frequency of 2000 Hz and a 3 dB bandwidth of 900 Hz. In the case of the single-formant signal, the bandpass filter serves only to generate the analytic signal while suppressing the effect of the source excitation further (compared to that achieved by pre-emphasis alone). The analytic signal is windowed by a Gaussian data window of length $(M + 1)$ samples. Since we are interested in studying the effect of data-window duration on the accuracy of the IF estimates, different window lengths corresponding to effective durations of 0.5,1 and 2 ms were applied. The DWD is computed as in Eq. (3) by using a zero-padded FFT of length 4096. This gives a frequency resolution of 1.95 Hz in the IF measurement. The IF is estimated at a given time instant by centering the DWD data window at that time instant and picking the frequency peak of the resulting DWD spectrum. Measurements are made at the locations referred to in Fig. 1. In the case of the noisy signal, the bandpass filter bandwidth was set at 450 Hz, which matches the filter bandwidth typically used in speech signal IF estimation. In this case IF measurements were made only in the closed phase over 10 samples at the centre and averaged further over 20 periods to estimate the root-mean-square (r.m.s.) error in the IF estimate.

4.1.3. Results

Plots of estimated IF versus time were obtained for the synthesised signals by calculating the IF at every sample. Fig. 3 shows plots of the estimated IF versus time as calculated at every sample, for the signal SS2 for the three different DWD data-window durations of 16 samples, 32 samples and 64 samples (corresponding to effective durations of 0.5,1 and 2 ms, respectively). A common characteristic of all the three plots is that the IF trajectory closely matches the true IF of the time-varying resonance at all transition rates (as represented by the dotted lines), except in regions where the source excitation is nonzero. Comparing the IF estimate plots on a sample-to-sample basis with the source excitation waveform of Fig. 2(a), we see that significant perturbations in the IF estimate occur in the region near the glottal closure instant and to a lesser extent near the glottal opening. In the pre-emphasised and bandpass filtered signal, the source excitation has a nonzero contribution in the regions near the glottal closure and opening instants causing abrupt phase changes in the resonance filter output. While the source excitation due to glottal closure causes a breakdown in the IF estimate, the source excitation contributed by the glottal opening in the waveform of Fig. 2(a) is less prominent and results in a minor ripple in the IF trajectory.

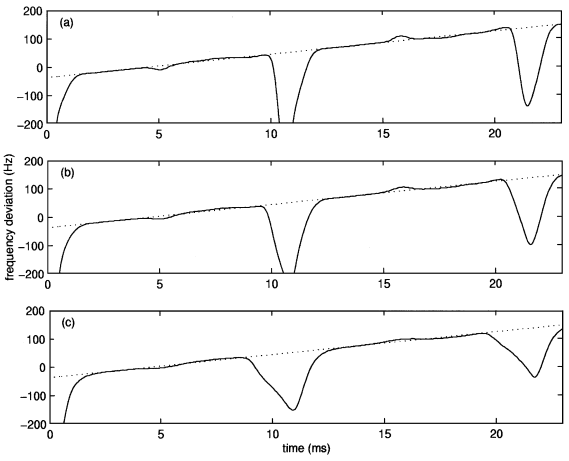

Fig. 3. IF estimates for linearly time-varying resonance, shown as the frequency deviation from a centre frequency of 2000 Hz, for different effective data-window durations. The dotted lines indicate the actual IF of the resonance. The DWD effective data-window lengths are (a) 0.5 ms; (b) 1 ms; (c) 2 ms.

Increasing the effective duration of the DWD data window from 0.5 to 2 ms causes a spreading of the perturbation due to the glottal closure since the larger data window is more likely to include portions of significant source excitation even when its centre is positioned well within the closed or open phases. When the source excitation is localised to a small region relative to the duration of the data window as is the case with the glottal opening, the

Table 1
Accuracy of the IF estimates for synthetic single-formant signal for different frequency transition rates, shown as frequency deviation (Hz) from the actual frequency for various locations in the glottal cycle (f_{i-16}: DWD data-window length = 16 samples; f_{i-32}: DWD data-window length = 32 samples; f_{i-64}: DWD data-window length = 64 samples)

Rate (Hz/ms)	Estimate	(i)	(iii)	(v)	(vi)	(viii)	(x)
0	f_{i-16}	+4	+4	+2	0	0	−10
	f_{i-32}	+4	+4	−4	0	0	−8
	f_{i-64}	+4	+2	−57	−4	0	−4
8	f_{i-16}	−3	0	+8	0	0	−9
	f_{i-32}	−3	−2	+6	0	0	−11
	f_{i-64}	−1	−2	−49	−5	0	−5
32	f_{i-16}	+6	−5	−12	+2	+2	−36
	f_{i-32}	+6	−5	+2	+2	+2	+18
	f_{i-64}	+4	−3	−23	−8	0	−6

longer data window results in smoother and more accurate estimates.

Table 1 presents some quantitative results for the accuracy of the IF estimate as measured by the estimation error at selected locations in the glottal cycle. We see that location (iii) which is approximately at the centre of the open phase and location (viii) at the centre of the closed phase, the IF estimate displays a high accuracy for all the three data windows and at all three transition rates. This is the case whenever that data window contains no component of the source excitation and the signal is well modelled by a free time-varying resonance. At the remaining locations on either side of the centre locations, the IF estimate is perturbed to various extents depending on the location and on the data-window duration. The values at location (v) close to the instant of glottal closure show the highest deviation in the case of the 2 ms window. In the case of the shorter windows, locations (v) and (x) near the closure and opening, respectively, are equally perturbed. In Table 2 are shown the r.m.s. error in the frequency estimate with the data window centred in the closed phase for the various SNR levels. We see that increasing the data-window length brings down the r.m.s. frequency estimation error. This can be attributed to the narrower frequency width of the DWD peak at larger window lengths. Looking at the overall accuracy we see that the 1 ms effective window length (32 samples) offers a good compromise between the smoothness of the IF estimates and maximising the duration of valid estimates in the glottal cycle.

Table 2
Root-mean-squared error (Hz) in IF estimates for synthetic single-formant signal for different SNR levels (f_{i-16}: DWD data-window length = 16 samples; f_{i-32}: DWD data-window length = 32 samples; f_{i-64}: DWD data-window length = 64 samples)

SNR (dB)	f_{i-16}	f_{i-32}	f_{i-64}
25	3.2	2.9	2.5
15	9.9	8.8	7.0
5	28.9	25.4	22.8

To put the above results in perspective, covariance-based LP analysis was applied to the same data by peak-picking of the LP spectrum. It was found that while a second-order LP analysis using a 3 ms data window located in the closed phase gave the best results, the estimate showed wide fluctuations with respect to the actual value in all the open-phase locations. Tenth-order LP analysis using the same data window gave smoother estimates but was significantly less accurate for the higher transition rate in all glottal cycle locations compared with the situation at the lower rates and compared with the IF estimate at all transition rates. These observations are in accordance with the results related to LP analyses presented in [7].

4.2. Multi-formant signals

4.2.1. Method of signal generation

Three-formant signals were generated by filtering the glottal-pulse derivative signal with a cascade of three second-order IIR filters each adjusted for a single formant. A set of signals is synthesised that incorporates the different cases of inter-formant frequency spacing, and fundamental frequency. Source–tract interaction is modelled by varying the frequencies and bandwidths of the formants synchronously with the source. That is, different parameters are used in each of the closed and open phases of the glottal cycle. The three synthetic signals generated are:

SM1: Vowel "ae", pitchperiod = 170 samples (fundamental frequency = 94.1 Hz), formant frequencies are as follows: closed phase (c.p.): $f_1 = 640$ Hz, $f_2 = 1720$ Hz, $f_3 = 2410$ Hz; open phase (o.p.): $f_1 = 700$ Hz, $f_2 = 1780$ Hz, $f_3 = 2470$ Hz; all formant bandwidths: in c.p. = 100 Hz; in o.p. = 150 Hz.

SM2: Vowel "ae", pitch period = 170 samples (fundamental frequency = 94.1 Hz), formant frequencies are as follows: c.p.: $f_1 = 640$ Hz, $f_2 = 1500$ Hz, $f_3 = 2400$ Hz; o.p.: $f_1 = 700$ Hz, $f_2 = 1560$ Hz, $f_3 = 2460$ Hz; all formant bandwidths: in c.p. = 100 Hz; in o.p. = 150 Hz.

SM3: Vowel "ae", pitch period = 120 samples (fundamental frequency = 134 Hz), formant frequencies are as follows: c.p.: $f_1 = 640$ Hz, $f_2 = 1500$ Hz, $f_3 = 2400$ Hz; o.p.: $f_1 = 700$ Hz, $f_2 = 1560$ Hz, $f_3 = 2460$ Hz; all formant bandwidths: in c.p. = 100 Hz; in o.p. = 150 Hz.

Signals SM1 and SM2 differ in the proximity of the second and third formants. The signal SM3 is similar to SM2 but with a lower pitch period.

4.2.2. Method of analysis

The multiformant signal was pre-emphasised and bandpass filtered about the closed-phase formant frequency of each formant using a bandwidth of 450 Hz. The rest of the processing was identical to the single-formant case except that the DWD data-window length was kept fixed at 32 samples.

4.2.3. Results

In Fig. 4 are shown the source excitation, the synthesised signal SM2 and its magnitude spectrum. Fig. 5 shows the IF estimate trajectory for each of the three formants of SM2. Comparing these with the source excitation waveform, we note the following (1) There are perturbations in the estimated IF in the vicinity of the instant of glottal

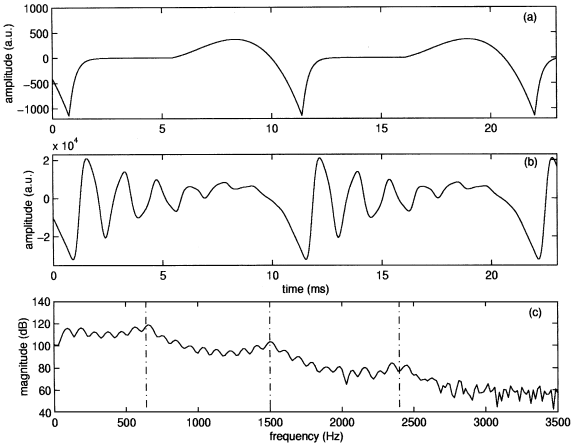

Fig. 4. (a) Source excitation at fundamental frequency = 94.1 Hz; (b) three-formant synthetic signal with time-varying formants; (c) magnitude spectrum of the signal showing the formants at 640, 1500 and 2400 Hz.

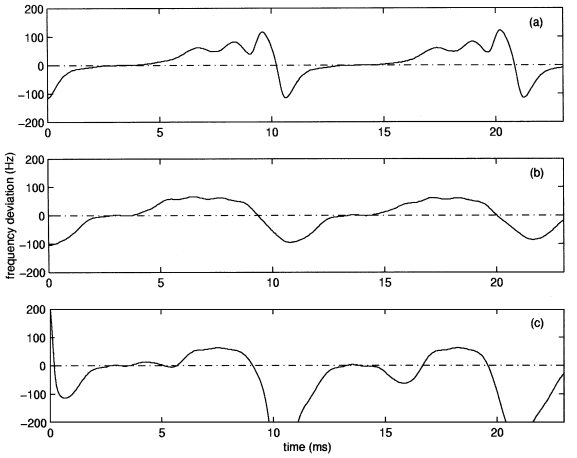

Fig. 5. IF estimates for each of the formants of the synthetic signal shown as the frequency deviation from a centre frequency (f_c): (a) formant 1, $f_c = 640$ Hz; (b) formant 2, $f_c = 1500$ Hz; (c) formant 3, $f_c = 2400$ Hz.

closure, and, to a lesser extent, at glottal opening as well. These are due to the abrupt phase changes that occur in the signal due to the significant source excitation present in these regions. The perturbations are more spread compared to the single-formant case due to the narrower bandwidth filter applied to the multi-formant signal. Processing the speech signal with the highpass and bandpass filters limits the effects of the source excitation to regions near the glottal closure and glottal opening instants, except in the case of the first formant. Here, we see a modulation in the IF estimate in the o.p. caused by the fact that the low centre-frequency bandpass filter does not fully suppress the source excitation in the o.p. (2) In regions well within the c.p. and the o.p., the IF estimate accurately reflects the frequency of the resonance in the case of all the three formants. There is a frequency spreading of the abrupt frequency changes at the glottal opening and closure which can be reduced by increasing the bandwidth of the bandpass filter, but of course, this is possible only to the extent that the neighbouring formants are adequately suppressed.

Table 3 presents some results on the accuracy of the IF estimates for all three synthetic signals. The frequency error at each time location in the glottal cycle is tabulated in terms of frequency deviation from the true value for each of the formants. The most accurate estimates occur at locations (iii) and (viii) (as expected). Comparing the values at these locations for all the three signals, we see that the best estimates are obtained for signal SM2. While the c.p. and o.p. estimates of IF for the second and third formants are equally accurate, that for the first formant in the o.p. deviates more due to its lower frequency and hence greater interaction with the source excitation. There is an increase in the error going to the case of the more closely spaced formants in SM1, due to the inability of the bandpass filter to isolate the second and third formants completely. In the case of SM3, where the pitch period is significantly less than in SM2, we see marked fluctuations in the error estimates compared to those of SM2. This is due to the fact that the DWD data window now includes more of the regions containing significant source excitation. It is possible to reduce the data-window duration to limit this effect but that would lead to an even greater loss in smoothness of the estimates in the open phase.

5. Experiments with natural speech

The DWD-based IF estimation method has been applied to several samples of voiced speech extracted from the TIMIT sentence database [2]. Voiced

Table 3
Accuracy of the IF estimates for synthetic three-formant signal, shown as frequency deviation from the actual frequency for various locations in the glottal cycle; (a) SM1; (b) SM2; (c) SM3

Location	(ii)	(iii)	(iv)	(vii)	(viii)	(ix)
(a)						
Δf_1	− 5	− 12	+ 21	− 5	+ 1	+ 1
Δf_2	− 22	+ 1	+ 15	− 5	− 1	+ 1
Δf_3	+ 32	− 11	− 36	+ 16	+ 2	− 29
(b)						
Δf_1	− 5	− 12	+ 19	− 3	+ 1	+ 1
Δf_2	− 11	− 1	+ 4	− 6	0	+ 4
Δf_3	− 22	+ 1	+ 38	− 5	0	− 5
(c)						
Δf_1	− 36	− 1	+ 25	− 1	− 1	+ 1
Δf_2	− 15	+ 4	− 21	− 59	− 6	+ 12
Δf_3	− 24	+ 5	+ 22	− 84	− 7	+ 22

speech segments, sampled at 16 kHz, corresponding to steady vowel sounds were analysed with the goal to observe formant frequency variations in the different phases of the glottal cycle.

We present here, in some detail, experimental results obtained with two samples of voiced speech from a male speaker (TIMIT database directory: test/dr2/mrgg0). Fig. 6(a) show the speech signal for the voiced sound "ae" over a duration of about 45 ms. The corresponding magnitude spectrum is shown in Fig. 6(b), with the approximate formant locations of the first three formants marked by dotted lines. The speech signal was processed as discussed in Section 3 with the bandpass filter centred at each formant location with a bandwidth of 450 Hz. Each of the filtered signals was analysed for the peak of the DWD spectrum with a sliding window of an effective data-window duration of 1 ms. Fig. 7 shows the estimated IF trajectories for each of the three formants. In order to be able to properly interpret the IF estimates, we need some knowledge of the source excitation waveform or at least, of the regions of each pitch period which contain significant excitation. The prediction error obtained from an 18th-order LP analysis using a 3 ms sliding data window is used for this purpose. The minima of the prediction error signal may be used to identify the time locations in the glottal

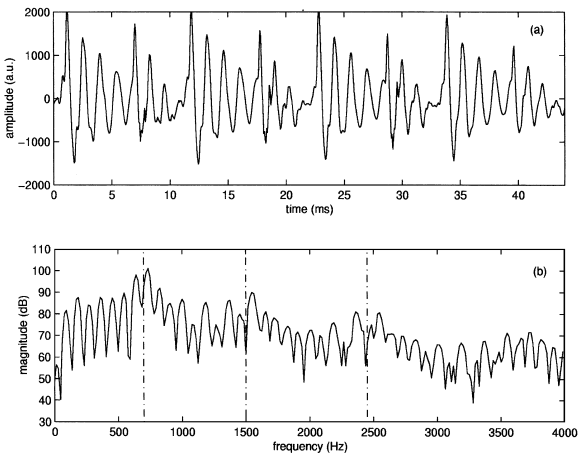

Fig. 6. Segment of male natural speech vowel "ae": (a) speech signal; (b) magnitude spectrum showing approximate formant locations.

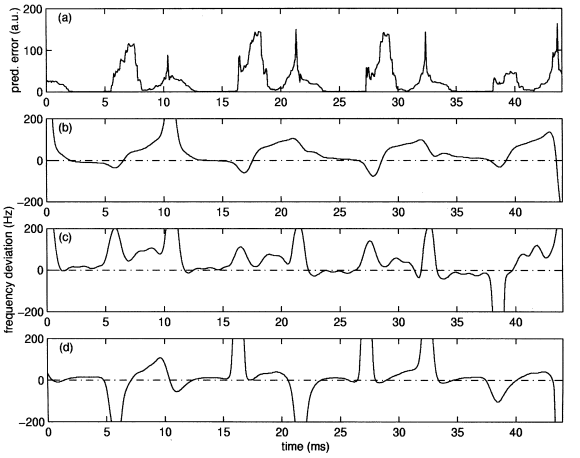

Fig. 7. IF estimates for the formants of the vowel "ae", shown as the frequency deviation from the centre frequency (f_c) of the corresponding bandpass filter: (a) prediction error; (b) IF estimate for formant 1, $f_c = 700$ Hz; (c) IF estimate for formant 2, $f_c = 1500$ Hz; (d) IF estimate for formant 3, $f_c = 2450$ Hz.

cycle at which the source excitation is negligible and the speech signal can be interpreted as the freely decaying oscillation of an all-pole filter. Only in these regions can the IF estimates be considered to be accurate. Fig. 7(a) shows the prediction error signal for the waveform of Fig. 6(a). We see that in this particular speech sample, the glottal closure as well as glottal opening are highly localised as indicated by near-zero prediction errors in regions of the closed phase as well as open-phase. From Figs. 7(b)–(d) we see that the overall nature of the IF estimates is very similar to the estimates obtained for the synthetic vowel. The perturbation in the IF estimates near the glottal opening is prominent indicating that the glottal opening must be more abrupt than is typical. Comparing the valid IF estimates in the c.p. and o.p., we note clear increases in the formant frequencies in the open phase for the first two formants. The IF estimate track for the second formant also indicates a gradual shift over time toward lower frequencies.

The second sample is the voiced sound "ow" extracted from a sentence spoken by the same speaker. In Fig. 8(a) is shown the waveform of a 25 ms segment of this speech sample. The corresponding magnitude spectrum is shown in Fig. 8(b), with

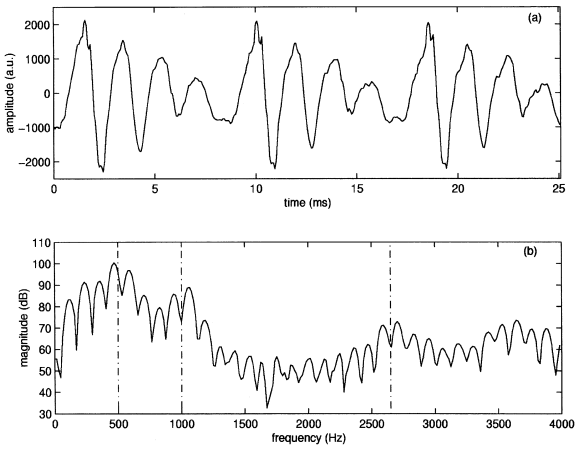

Fig. 8. Segment of male natural speech vowel "ow": (a) speech signal; (b) magnitude spectrum showing approximate formant locations.

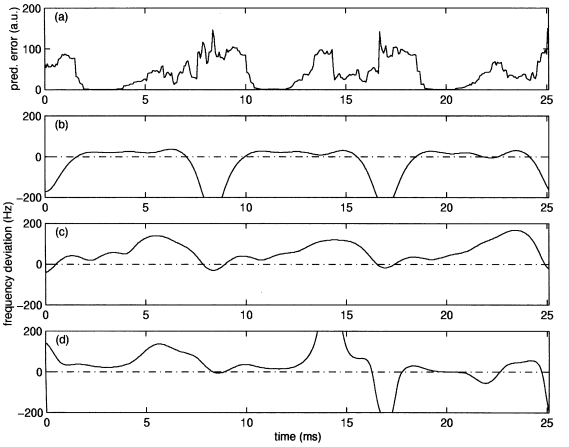

Fig. 9. IF estimates the formants of the vowel "ow", shown as the frequency deviation from the centre frequency (f_c) of the corresponding bandpass filter: (a) prediction error; (b) IF estimate for formant 1, $f_c = 500$ Hz; (c) IF estimate for formant 2, $f_c = 1000$ Hz; (d) IF estimate for formant 3, $f_c = 2650$ Hz.

the approximate formant locations of the first three formants marked by dotted lines. The first two formants are relatively closely spaced at 500 and 1000 Hz, respectively. It is necessary to use a lower bandwidth for the bandpass filter to isolate the first two formants. A bandwidth of 350 Hz was used for both formants. In Fig. 9(a) is shown the prediction error obtained as described earlier. We see that while the prediction error goes to zero in the closed phase indicating complete closure, the error is relatively high in the open phase. Figs. 9(b)–(d) show the estimated IF trajectories for each of the three formants. The first and third formants were analysed with a 1 ms effective DWD data window. In the case of the second formant it was found necessary to increase the data window to 2 ms to obtain smooth estimates in the open phase. This may be attributed to the presence of nonzero source excitation in the open phase.

Summarising our observations from the analyses of the first three formants of several natural speech vowels, we find that for voiced sounds in which (i) the formant frequencies are well separated, and (ii) there are distinct portions in the pitch period, corresponding to the closed and open phases in which the source excitation can be localised to a narrow region by filtering, it is possible to observe the variations in vocal-tract resonance frequencies in the different phases of the glottal cycle by IF estimation. A study of several segments of vowel sounds taken from the same male speaker referred to earlier, has shown consistent increases across pitch periods and across occurrences in the open-phase frequencies of the first two formants of the sounds "ae", "ey","iy" and "ow" (examples of their occurrence are in the words "that","save","me" and "don't", respectively). The second formant is particularly consistent in showing abrupt increases in the open-phase value from the closed-phase value by 60–100 Hz. Similar results have been reported for other vowels in [11] based on frequency estimates from covariance-based LP analyses. The third formant sometimes shows differences in closed- and open-phase frequency but due to its relatively low strength, the IF estimate is more likely to be corrupted by the influence of neighbouring formants and by the source excitation. The formant frequency variations within a pitch period are often accompanied by a gradual time shift in the formant frequencies over successive pitch periods, an effect which is captured well by the instantaneous frequency estimator.

6. Conclusions

The problem addressed here was the tracking formant frequency variations in voiced speech,

particularly variations due to source–tract interaction. Since the problem involves the high-resolution tracking of time-varying frequencies, the use of instantaneous frequency estimation methods was motivated. Instantaneous frequency estimation based on the Wigner distribution time–frequency representation was described. While such methods can provide accurate estimates of the frequency of the free resonance of a time-varying system, application to speech analysis necessitates considerable pre-processing of the speech signal as well as a careful interpretation of the resulting estimates. While bandpass filtering of the speech is necessary to isolate the individual resonances, added high-pass filtering is essential to minimise the effects of source excitation in the analysis of lower frequency formants. The bandpass filtering leads to a time-domain smoothing of the signal and hence of the instantaneous frequency. The presence of source excitation results in inaccurate IF estimates due to the resulting abrupt changes in the signal phase. The sensitivity to source excitation as well to any additive noise in the signal can be minimised by increasing the effective duration of the DWD data window. This however leads to a loss in the time resolution of the IF estimate if the time variation of frequency is not linear. So while instantaneous frequency estimation is conceptually well suited for the problem of tracking time-varying resonances, adapting it for implementation on speech signals to estimate the frequencies of a time-varying vocal-tract system is subject to limitations in the achievable time resolution. Such limitations may make it only comparable (rather than superior) to a quasi-stationary method in most cases.

As long as the formants are well separated in frequency and there is no significant level of source excitation present, IF estimation as described here can be used to estimate the variation in the formant frequencies in the distinct phases of the glottal cycle in voiced speech with the proper choice of filter bandwidth and DWD data-window length.

Acknowledgements

The authors would like to thank an anonymous reviewer for valuable comments and for making them aware of Reference [3].

References

[1] D.G. Childers, Measuring and modeling vocal source–tract interaction, IEEE Trans. Biomed. Eng. BE-41 (1994) 663–671.
[2] DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus, NIST, 1990.
[3] R. Kumaresan, A. Rao, Model-based approach to envelope and positive instantaneous frequency of signals with speech applications, J. Acoust. Soc. Am. 105 (3) (1999) 1912–1924.
[4] P. Maragos, J. Kaiser, T. Quatieri, Energy separation in signal modulations with application to speech analysis, IEEE Trans. Signal Process. SP-41 (1993) 3024–3051.
[5] P. Rao, A robust method for the estimation of formant frequency modulation in speech signals, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Atlanta, 1996.
[6] P. Rao, Estimation of instantaneous frequency using the discrete-Wigner distribution, Electron. Lett. 26-4 (1990) 246–248.
[7] R. Smits, Accuracy of quasistationary analysis of highly dynamic speech signals, J. Acoust. Soc. Am. 96 (6) (1994) 3401–3415.
[8] D. Vakman, On the analytic signal, the Teager–Kaiser energy algorithm, and other methods for defining amplitude and frequency, IEEE Trans. Signal Process. SP-44 (1996) 791–797.
[9] R. Veldhuis, A computationally efficient alternative for the LF model and its perceptual evaluation, J. Acoust. Soc. Am. 103 (1) (1997) 566–571.
[10] W. Wokurek, F. Hlawatsch, G. Kubin, Wigner distribution analysis of speech signals, Proceedings of the International Conference on Digital Signal Processing, Italy, 1987.
[11] B. Yegnanarayana, R.N.J. Veldhuis, Extraction of vocal-tract system characteristics from speech signals, IEEE Trans. Speech Audio Process. 6 (4) (1998) 313–327.