Extraction of Vocal-Tract System Characteristics from Speech Signals

B. Yegnanarayana, *Senior Member, IEEE,* and Raymond N. J. Veldhuis

***Abstract*— We propose methods to track natural variations in the characteristics of the vocal-tract system from speech signals. We are especially interested in the cases where these characteristics vary over time, as happens in dynamic sounds such as consonant-vowel transitions. We show that the selection of appropriate analysis segments is crucial in these methods, and we propose a selection based on estimated instants of significant excitation. These instants are obtained by a method based on the average group-delay property of minimum-phase signals. In voiced speech, they correspond to the instants of glottal closure. The vocal-tract system is characterized by its formant parameters, which are extracted from the analysis segments. Because the segments are always at the same relative position in each pitch period, in voiced speech the extracted formants are consistent across successive pitch periods. We demonstrate the results of the analysis for several difficult cases of speech signals.**

***Index Terms*—Formant analysis, speech analysis.**

I. Introduction

THE OBJECTIVE of this paper is to propose methods to track natural variations in the characteristics of the vocal-tract system from speech signals. The information in these variations may be crucial for applications in speech recognition. Understanding and modeling these variations is also useful for speech synthesis.

The shape of the vocal-tract system is determined by the positions of the articulators. The vocal-tract shape is difficult to derive from the speech signal [1], [2]. Therefore, we use the formant parameters, which are commonly used in speech analysis and synthesis, to characterize the vocal-tract system. The formants are the free resonances of the vocal-tract system. A formant is described by three parameters: i) the formant frequency, ii) the formant bandwidth or, equivalently, the damping, and iii) the formant amplitude.

The resonances of the vocal tract, and thus the formant parameters, vary in time in two distinct ways. First, the shape of the vocal tract varies during the production of speech, due to the movement of articulators. It is these variations that we find interesting in many applications. They are usually slow during speech production, except during the transitions in the production of consonant-vowel units. Second, the formant parameters vary within one pitch period, even though the articulators themselves do not move [3]–[5]. This is because the vocal folds oscillate between an open and a closed phase, thus, during each glottal cycle the system characteristics change. During the closed phase the vocal tract is closed at one end and the speech signal is mainly due to free resonances, but during the open phase the trachea, the vocal folds, and the vocal tract are acoustically coupled, and this coupling will change the free resonances. Actually, the situation is still more complicated. During the open phase, the air flow through the vocal folds increases initially and subsequently decreases as a function of time. The relation between the acoustic pressure over the vocal folds and the air flow is in general nonlinear [3], [6]–[9]. Hence, the characteristics of the system during the open phase are not constant, but signal dependent. Fig. 1 shows the waveform $s(t)$ of a few periods of a sustained vowel /a/, the corresponding glottal airflow $g(t)$, consisting of a sequence of glottal pulses, and the time derivative of the glottal airflow dg/dt. Both $g(t)$ and dg/dt have been derived from the acoustic signal by means of an inverse-filtering technique similar to the one described in [10]. The assumed open and closed phases are marked in the figure by a 'C' and an 'O,' respectively. Compared with the signal in the closed phase, the signal in the open phase shows a higher damping of the resonances. We can see that the shape of the time derivative of the glottal airflow is reflected in the speech signal. Because of the changes within a pitch period, it is necessary to determine the formant parameters separately for each of the open and closed phase regions. In practice the closed phase can be very short to the extent that it may vanish completely. For example, in high-pitched (e.g., female) voices the vocal folds have been observed to start opening directly after closure [5]. Closure may also not be complete, in which case some leakage occurs [11].

We will only consider voiced speech and we will adopt the well-known source-filter model [6], [7], [12] for the analysis of the speech signal. The estimation problem is then a model-parameter estimation problem. The source-filter model consists of a source that generates a sequence of glottal pulses, modeling the glottal air flow. This is input to a filter that models the vocal-tract system, which includes a differentiation operator that models the radiation at the lips. In this model, the radiation operator and the vocal-tract filter are interchanged and the radiation operator operates directly on the glottal pulses. The combination of the glottal source and the radiation operator is replaced by a single source producing differentiated glottal pulses. This approach is similar to the one followed in [13]. Glottal closure in voiced speech is generally abrupt. Therefore, the presence of the differentiator results in a strong

Manuscript received May 14, 1996; revised September 12, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

B. Yegnanarayana is with the Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India.

R. N. J. Veldhuis is with IPO-Center for Research on User-System Interaction, 5600 MB Eindhoven, The Netherlands (e-mail: veldhuis@ipo.tue.nl).

Publisher Item Identifier S 1063-6676(98)04219-9.

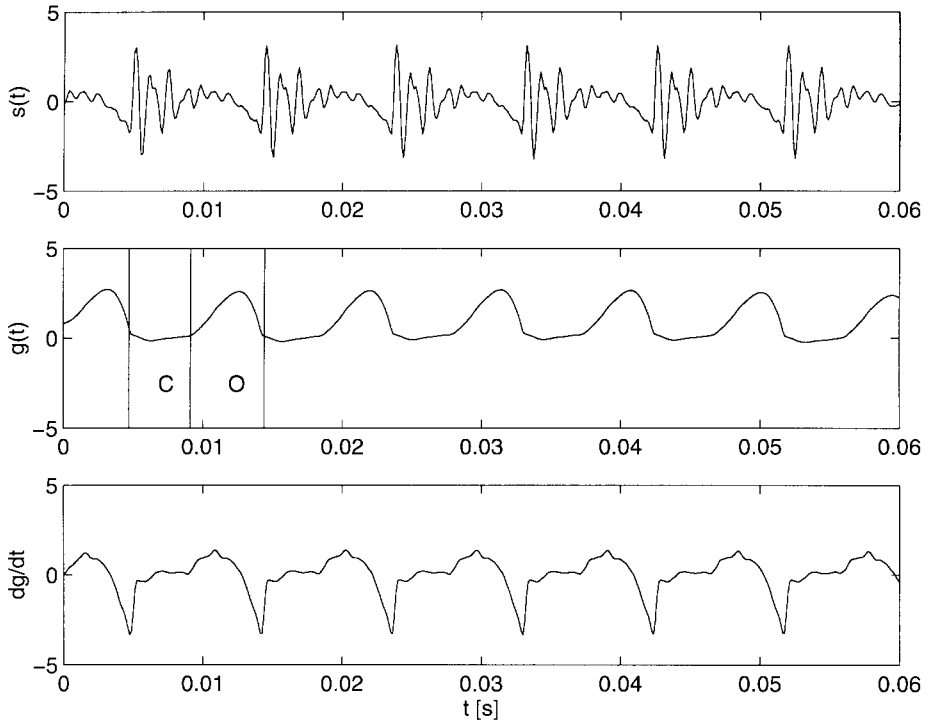

Fig. 1. Male vowel /a/. Top: waveform $s(t)$; sampling frequency 8 kHz. Middle: glottal pulse $g(t)$; the closed phase (C) and the open phase (O) are indicated. Bottom: time derivative of glottal pulse dg/dt.

spikelike excitation of the vocal-tract filter just preceding the glottal closure. This explains why this is an instant of significant excitation.

Let us first consider the closed phase. Clearly, the input signal vanishes in this situation, so that the speech signal consists only of the free resonances of the vocal-tract system. This behavior can be modeled by an all-pole filter, except in the case of nasals or nasalized vowels, where a pole-zero filter may be more appropriate. A problem is that the closed phase can be short or in some cases nonexistent.

In the open phase, the situation is more complex than in the closed phase. First, the time derivative of the glottal pulse forms a nonzero input to the vocal-tract filter. Second, the system includes not only the vocal tract, but also the trachea and the coupling of the trachea and the vocal tract is time varying, due to the vocal-fold motion. Third, the overall system shows some nonlinear behavior. The presence of the subglottal tract including the trachea has several effects [3]–[5], [11], [14]. It will increase the damping, shift the resonance frequencies, and may introduce additional poles and zeros. Furthermore, it is difficult to analyze the nonlinear behavior because good models for it do not exist. In this study, we assume that the nonlinear effects are not significant. Thus, the open-phase model consists of an unknown source exciting a time-varying pole-zero filter.

If there is a closed phase, the estimation of formant parameters in it is relatively simple, because the vocal-tract system is stationary and has zero input. The unknown source and the time-varying characteristics in the open phase make it much more difficult to estimate the free resonances than in the closed phase. We may be able to compensate for the unknown source by means of a preemphasis filter. However, the formant parameters obtained from an analysis frame in the open phase will always be some kind of time averages. It is clear that one has to be careful with the interpretations of the formant parameters that were estimated in the open phase.

Current methods of speech analysis typically use subsequent blocks of 10–20 ms of data to estimate the characteristics of the vocal-tract system in the interval of the block. Block processing smears the information within the analysis frame and gives an estimate of the spectrum corresponding to some averaged behavior. The size and shape of the analysis frame also affects the estimated spectral characteristics, as does the position of the analysis frame with respect to the signal. In addition, it is well known that the fundamental frequency significantly influences the spectrum if the analysis frame contains more than one pitch period. For example, for speech with a high fundamental frequency the shapes of the short-time spectral envelope and the derived linear-prediction spectrum depend on the voice harmonics [15]. On the other hand, for short (less than one pitch period) data records, the performance of high-resolution techniques such as covariance linear-prediction analysis depends critically on the position of the analysis frame within a pitch period because the signal properties can be significantly different in different nearby regions, such as the closed and open phases [16]. These varying properties can produce significantly different estimates of the formant parameters and may mask the natural variations. A way to deal with some of the problems of estimating formant parameters was presented in [17]. In this block-based method the group-delay spectrum was used to determine the formant frequencies. Interestingly, the present paper proposes a completely different

use of the group delay as a solution to the same problems. Pitch-synchronous analysis is reported to give better results than pitch-asynchronous analysis [10], [18], [19]. However, in most cases it is difficult to determine the pitch-synchronous instants automatically from the speech signal. In [18], for instance, an electroglottograph signal was used to determine the closed phases. In [10] the local minima of the normalized prediction error were used for this purpose, and [19] uses the peaks in the linear prediction residual to identify the instants of glottal closure.

The main objective of this study is to track the small natural variations in the formant parameters during the production of speech. To overcome some of the problems of the block-processing approach, we propose a pitch-synchronous analysis method that is based on the knowledge of the instants of significant excitation of the vocal-tract system. Recently, a method was proposed to determine such instants from speech signals [20], [21]. In voiced speech, these instants typically just precede glottal closure. Knowledge of these instants enables us to choose the position and the size of the analysis frame within a pitch period in such a way that we can avoid smearing of features and fluctuation of the estimated parameters. In particular, the position of the analysis frame in consecutive pitch periods will be consistent with respect to the instant of significant excitation. In addition, we aim to choose the size of the analysis frame such that the frame is either mainly in the closed or mainly in the open phase. There is no possibility of multiple pitch periods in one analysis frame corrupting the estimation of parameters, because the frame size is always smaller than a pitch period. The use of an analysis frame that is short compared to the length of a pitch period and that is positioned directly after the instant of excitation enables the tracking of heavily damped formants that cannot be observed when larger analysis frames are used. In the case of a higher fundamental frequency, the analysis frame may become too short for the reliable extraction of the parameters. As a solution, we have developed a pitch-synchronous averaging technique called the *multicycle covariance method*, which averages covariance estimates over a number of consecutive pitch periods. It will be shown that, with the method presented here, it is possible to derive the temporal variations of the formant parameters accurately.

The first step in this method is to identify the distinct phases of speech production in voiced speech. The detection of the instants of significant excitation, which will enable us to isolate these phases, is described briefly in Section II. With the knowledge of these instants, it is possible to analyze the characteristics of the vocal-tract system in a pitch-synchronous manner. The analysis methods are described in Section III. In Section IV, the effects of size and position of the analysis frame on the estimated formant parameters are discussed and the method is compared with pitch-asynchronous methods. The results of the analysis for different types of speech segments are discussed in Section V in order to demonstrate the ability of the proposed instants-based approach to extract dynamic characteristics of the vocal-tract system. Section VI presents conclusions.

II. Extraction of Instants of Significant Excitation

Recently, a method was proposed for determining the instants of significant excitation in speech signals [20], [21]. The method is based on the assumption that the speech signal contains delayed versions of minimum-phase impulse responses, each of which is the response to a significant excitation. A minimum-phase impulse response has zero average group delay. Consequently, if an analysis window contains a single major excitation, the average group delay of the signal in that window will be equal to the position of the major excitation with respect to the beginning of the window. The algorithm computes the average group delay as a function of time and marks the location of positive-going zero crossings as instants of significant excitation. The method is summarized in the following paragraphs.

In order to estimate the group delay as a function of time the following two functions of a time index m and frequency index n are derived from the time signal $x(m)$ as follows:

$$X(m,n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} w_a(k)x(k-mS)e^{-ikn(2\pi/N)}, \quad m \in \mathbb{Z}, \qquad n = 0, \cdots, N-1 \tag{1}$$

$$Y(m,n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} kw_a(k)x(k-mS)e^{-ikn(2\pi/N)}, \quad m \in \mathbb{Z}, \qquad n = 0, \cdots, N-1. \tag{2}$$

The function $X(m,n)$ is a discretized short-time Fourier transform at time mS/f_s and at frequency $f_s n/N$. Here S is the window shift, f_s the sampling frequency, and $\{w_a(k)\}_{k=0,\cdots,N-1}$ is a real-valued analysis window function. The symbol $\mathbb{Z}$ denotes the set of integers. The function $Y(m,n)$ can be seen as the derivative of the short-time Fourier transform with respect to frequency. In the present application, $S = 1$ and the analysis window is of the Hanning type. The window length is not very critical, although a window size of twice the length of an average pitch period is recommended.

For each window starting at mS/f_s the group-delay function at frequency $f_s n/N$ is computed [22] as

$$\tau(m,n) = \frac{X_R(m,n)Y_R(m,n) + X_I(m,n)Y_I(m,n)}{X_R^2(m,n) + X_I^2(m,n)} \quad m \in \mathbb{Z}, \qquad n = 0, \cdots, N-1 \tag{3}$$

with $X_R(m,n) + iX_I(m,n) = X(m,n), Y_R(m,n) + iY_I(m,n) = Y(m,n)$. For a given time index m the group-delay function $\tau(m,n)$ is first smoothed in the frequency domain, using a median filter, and subsequently its average over the frequency index n is computed. The resulting function is the phase-slope function $\psi(m)$. Instants where the phase-slope function makes a positive zero crossing are identified as instants of significant excitation. Here, significant excitation refers mainly to the instant of glottal closure in voiced speech, although the method also gives the instants at the onset of other significant events like bursts, release-of-stop sounds, and secondary excitations caused by glottal opening in voiced speech. In the unvoiced, silent, and aspirated regions, the instants are randomly positioned. Typically, the instants in

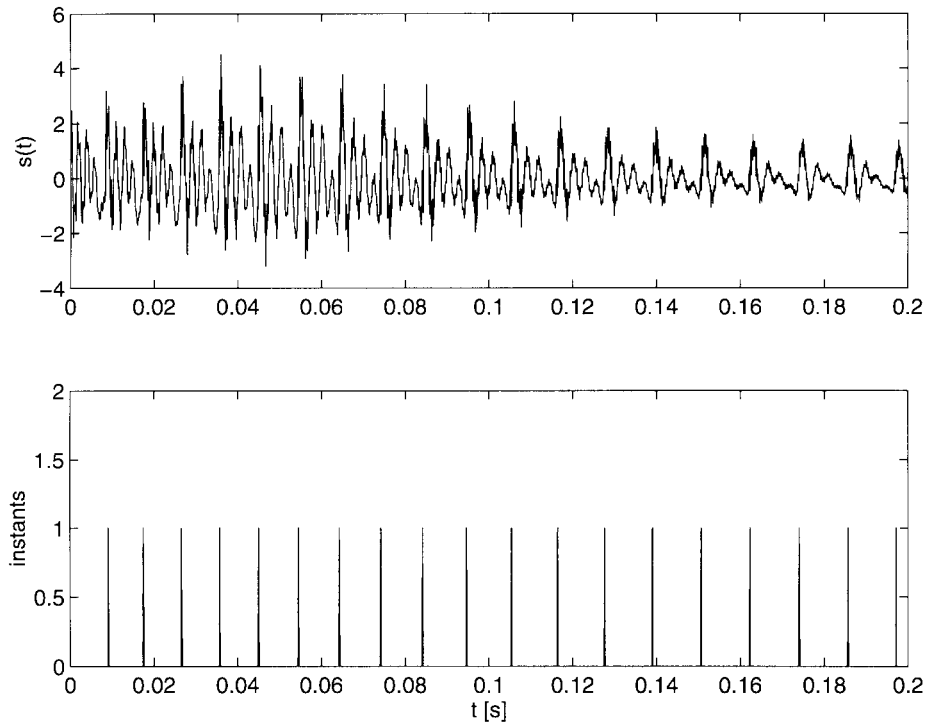

Fig. 2. Male diphthong /eI/. Top: waveform $s(t)$; sampling frequency, 8 kHz. Bottom: instants of excitation.

the voiced regions can be distinguished from those in the unvoiced regions by their quasiperiodic nature. The instants in the unvoiced and silence regions can be distinguished, if necessary, using the local signal power and the average spacing between the instants. The results improve if the analysis is done on a preemphasized speech signal $s(n) - \alpha s(n-1)$, with $0.8 < \alpha < 1$.

Fig. 2 shows a voiced speech signal and impulses positioned at the estimated instants of significant excitation. From the figure it is clear that the moments of acoustic response of the vocal tract to the voice source are close to the estimated instants of significant excitation.

III. Extraction of Formant Parameters

A. Formant Parameters

In this section, we discuss methods to determine the formant parameters of voiced speech using the knowledge of the instants of significant excitation, which coincide with the instants of glottal closure. The analysis is performed in order to characterize the time-varying behavior of the vocal tract due to the movement of the articulators and to distinguish its characteristics in closed and open glottal phases. In order to track the time-varying characteristics of the vocal-tract system, it is necessary to consider short analysis frames (less than one pitch period) of speech. These analysis frames are chosen around the instants of significant excitation.

For the analysis, it is important to decide how the desired parameters are to be extracted. A straightforward short-time spectral envelope will not be useful, as the short duration of the signal will not give enough spectral resolution. An all-pole model or a pole-zero model fit for the data segment would be useful to bring out the resonance and antiresonance characteristics of the vocal-tract system. For a given analysis frames first an appropriate all-pole or pole-zero model is determined and then the frequencies of the complex poles of the model are extracted. Pole frequencies below 200 Hz are considered spurious and are ignored [19]. Although our interest is in the tracking of the formant frequencies F_k, we will also study the behavior of the formant bandwidths B_k.

Each formant is a free resonance of the vocal-tract system, thus the corresponding time signal can be written as a sum of complex resonances, as follows:

$$r(n) = \sum_{k=1}^{p} A_k \rho_k^n e^{i\theta_k n} = \sum_{l=1}^{p/2} \rho_l^n (A_l e^{i\theta_l n} + \overline{A}_l e^{-i\theta_l n}). \quad (4)$$

Here, n is the time index, p equals twice the number of formants with frequencies below $f_s/2, k$ is the index of the particular formant, $\theta_k, -\pi < \theta \leq \pi$, is the normalized formant frequency, $\rho_k, 0 \leq \rho < 1$, determines formant damping, and A_k is the complex formant amplitude. The right-hand side of the equation holds because $r(n)$ is real valued and, therefore, the formant resonances in (4) occur in complex-conjugate pairs. The actual formant frequency F_k and bandwidth B_k values in Hz are given [16] by

$$F_k = \frac{f_s}{2\pi}\,\theta_k \quad (5)$$

$$B_k = -\frac{f_s}{\pi}\,\ln(\rho_k). \quad (6)$$

The z-transform of the time signal in (4), assuming a half-infinite sequence starting at $n = 0$, is given by

$$R(z) = \sum_{k=1}^{p} \frac{A_k}{1 - \rho_k e^{i\theta_k} z^{-1}} \quad (7)$$

$$= \frac{b_0 + b_1 z^{-1} + \cdots + b_{p-1} z^{-(p-1)}}{a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}}. \quad (8)$$

Note that, due to the arbitrary formant amplitudes A_k, $R(z)$ is not necessarily the z-transform of an all-pole transfer function. However, $R(z)$ can be regarded as the z-transform of the impulse response of an infinite impulse response filter. We will make use of this fact in Subsection III-B. The formant frequencies F_k and bandwidths B_k can be derived from the roots $\rho_k e^{i\theta_k}$ of the prediction polynomial

$$A(z) = a_0 + a_1 z^{-1} + \cdots + a_p z^{-p} \quad (9)$$

by using (5) and (6). The formant amplitudes, if required, can be obtained by using the b_k to obtain a representation as in (7).

B. Formant Analysis Methods

We make a distinction between formant parameters obtained from analysis frames before and after the instant of excitation, which we will call pre- and postexcitation parameters, respectively. If there is a distinct closed phase and if the analysis frame fits in it, then the postexcitation parameters represent the vocal-tract characteristics without being influenced by the glottal waveform, the subglottal tract, and system nonlinearities. In the derivation of the analysis methods, we assume that the postexcitation analysis frame lies in the closed phase, although we realize that this may not be the case in practice, and that the preexcitation analysis frame lies in the open phase. Correct positioning of the analysis frames in the closed or open phases requires the availability of an electroglottograph signal [18] or automatic inverse filtering, e.g., [5], [10], on running speech, for which no reliable methods are known to the authors. The minimum number of samples required in an analysis frame depends on the number of formants and on the method used to compute the formant parameters. It is preferable to use as many samples as possible, but the larger the number, the more likely it becomes that the vocal-tract system characteristics will change within the analysis frame.

First we consider a postexcitation analysis frame, which is assumed to be in the closed phase and, therefore, is modeled as a sequence with a z-transform as in (7) or (8). Our starting point is the difference equation corresponding to (8)

$$\sum_{k=0}^{p-1} b_k e(n-k) = \sum_{l=0}^{p} a_l s(n-l), \qquad a_0 = 1. \quad (10)$$

Here, $e(n)$ is an assumed excitation signal, $s(n)$ is the speech signal, and n is the time index. The analysis frame contains the samples $s(0), \cdots, s(N-1)$. Since (8) can be seen as an infinite impulse response, we define $e(0) = 1$ and $e(k) = 0, k = 1, \cdots, N-1$. In addition, $s(n) = 0$ for $n < 0$.

The problem is to estimate the a_k and b_k from the $s(0), \cdots, s(N-1)$. The standard approach to this problem is the Prony method [16], which minimizes

$$\sum_{n=p}^{N-1} \left(s(n) + \sum_{l=1}^{p} a_l s(n-l) - \sum_{k=0}^{p-1} b_k e(n-k) \right)^2 \quad (11)$$

as a function of the $a_1, \cdots, a_p$ and the $b_0, \cdots, b_{p-1}$. As a result, the $a_1, \cdots, a_p$ are obtained from solving

$$\boldsymbol{C}\boldsymbol{a} = -\boldsymbol{c} \quad (12)$$

where $\boldsymbol{C}$ is the $p \times p$ covariance matrix with elements

$$c_{kl} = \sum_{n=p}^{N-1} s(n-k)s(n-l), \qquad k, l = 1, \cdots, p \quad (13)$$

and the vector $\boldsymbol{c}$ has elements

$$c_k = \sum_{n=p}^{N-1} s(n-k)s(n), \qquad k = 1, \cdots, p. \quad (14)$$

This part of the Prony method is identical to the covariance method for estimating linear prediction coefficients [16]. After the $a_1, \cdots, a_p$ have been computed, the $b_0, \cdots, b_{p-1}$ follow from

$$b_k = s(k) + \sum_{l=1}^{p} a_l s(k-l). \quad (15)$$

Since the way of computing the $a_1, \cdots, a_p$ is identical to the covariance method, this method has the same drawback as the covariance method; namely, that $a_1, \cdots, a_p$ may correspond to an unstable filter.

The formant frequencies F_k and the bandwidths B_k can be derived from the roots $\rho_k e^{i\theta_k}$ of the prediction polynomial $a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}$, by using (5) and (6). Roots with a magnitude ρ_k below a certain threshold, say 0.8, and with a normalized absolute frequency $|\theta_k|$ also below a certain threshold, say corresponding to a frequency of 200 Hz, are assumed not to be due to formant resonances and are omitted [19]. The formant amplitudes, if required, can be obtained by using the b_k to obtain a representation as in (7).

If we want to estimate the preexcitation formant parameters we can follow the same approach, but have to take into account the influence of the excitation signal on the resonance system. The signal $e(n)$ in (10) can in this case not be assumed to vanish for $n > 0$. We may choose to ignore this influence or perform the analysis on the preemphasized signal $s(n) - s(n-1)$. This is a highpass version of the signal in which the influence of the time-derivative of the glottal pulse, which has a frequency roll-off approximately between 6 and 12 dB per octave above a cut-off frequency of approximately 100 Hz (e.g., [3], [7], [23]), has been reduced.

High-pitched (e.g., female) or noisy voices may cause some additional problems. Due to a higher fundamental frequency, the closed phase may be too short to obtain reliable estimates for the formant frequencies, especially when the speech is noisy. For example, if the fundamental frequency equals 200 Hz, which is not very high, and the open quotient is 0.6, then the closed phase will be as small as 2 ms. Increasing the frame length may improve the consistency of the measurement, but not its reliability, since the postexcitation analysis frame then may contain a part of the open phase. We can improve the results by using the samples of a limited number of consecutive postexcitation analysis frames. This is reasonable, since as the fundamental frequency increases, the fluctuations of the

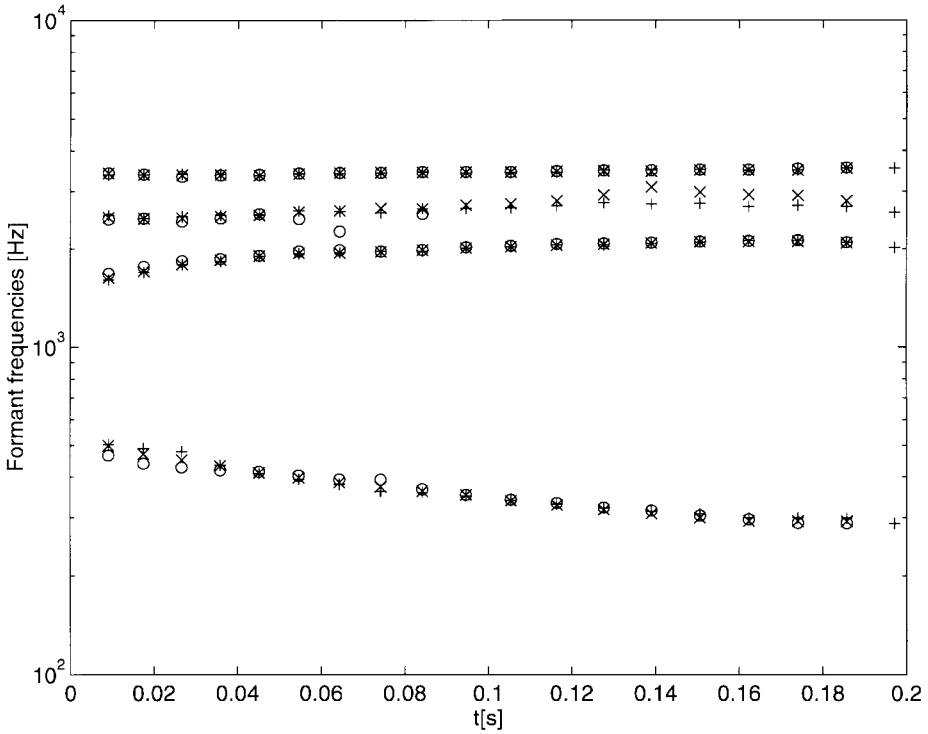

Fig. 3. Postexcitation formant frequency tracks of male diphthong /eI/. Order of prediction: 9. Symbols "+": frame length, 2.5 ms. Symbols "×": frame length, 5 ms. Symbols "○": frame length, 10 ms.

formant parameters become slower with respect to the pitch period. However, at higher formant transition rates a flattening-off effect may occur in the estimated formant frequencies as well as an increase of estimated formant bandwidth. The flattening-off effect was observed in [24], where it was studied as a function of the analysis-window length and the formant transition rate. When we combine consecutive postexcitation analysis frames, the prediction coefficients $a_1, \cdots, a_p$ are solved from

$$\left(\sum_{k=0}^{K-1} \boldsymbol{C}_k\right)\boldsymbol{a} = -\left(\sum_{k=0}^{K-1} \boldsymbol{c}_k\right) \tag{16}$$

where the $\boldsymbol{C}_k, k = 0, \cdots, K-1$, and the $\boldsymbol{c}_k, k = 0, \cdots, K-1$, are the $p \times p$ covariance matrices (12) and the covariance vectors (13) for K consecutive pitch periods. We will call this method the multicycle covariance method. The concept of pitch-synchronous averaging for formant estimation was also used and motivated in [4], but there the averaging was over successive waveforms. This may lead to partial suppression of formants due to small phase differences between the successive pitch periods. These phase differences, in their turn, can be the consequence of a fundamental frequency that is not a divisor of the sampling frequency.

IV. Performance Analysis

In this section, we analyze the effects of size and position of the analysis frame on the estimated formant frequencies and we show that analysis frames synchronized with the instants of significant excitation give consistently better results than uniformly spaced analysis frames.

The formant frequencies are derived from the arguments of the roots of the prediction polynomial (9) by using (5). The order of prediction was 9, which allows a combination of at most eight complex poles for four formants and one real pole to model the spectral behavior of the glottal waveform [15], [16], [25], [26]. Only roots with a magnitude greater than 0.8, corresponding to a bandwidth of 570 Hz and a frequency greater than 200 Hz, have been taken into account [19]. These parameter settings will be used throughout the remainder of the paper. The formant tracks are also presented as raw data without any form of smoothing throughout the paper. Suitable smoothing algorithms can be found in [16].

A. Effects of Size and Position of the Analysis Frame

As a first example we use the diphthong /eI/, as in "laid," produced by a male speaker. Fig. 2 shows the speech segment along with the extracted instants of significant excitation. The sampling frequency was 8 kHz.

Fig. 3 shows postexcitation formant frequencies, measured with analysis frame sizes of 2.5 ms, 5.0 ms, and 10 ms, denoted with plot symbols "+," "×," and "○," respectively. There are only small differences between the formant tracks, except that the third formant at about 2500 Hz partially disappears when the 10-ms frames are used. The reason for this phenomenon, as will become clear when formant bandwidth is discussed, is the higher bandwidth, and therefore higher damping, of this formant. Consequently, the third formant is only significantly present in the first few samples of the frame and its relative influence becomes weaker when the frame length increases. It disappears after a few frames, because the bandwidth of the third formant increases with time.

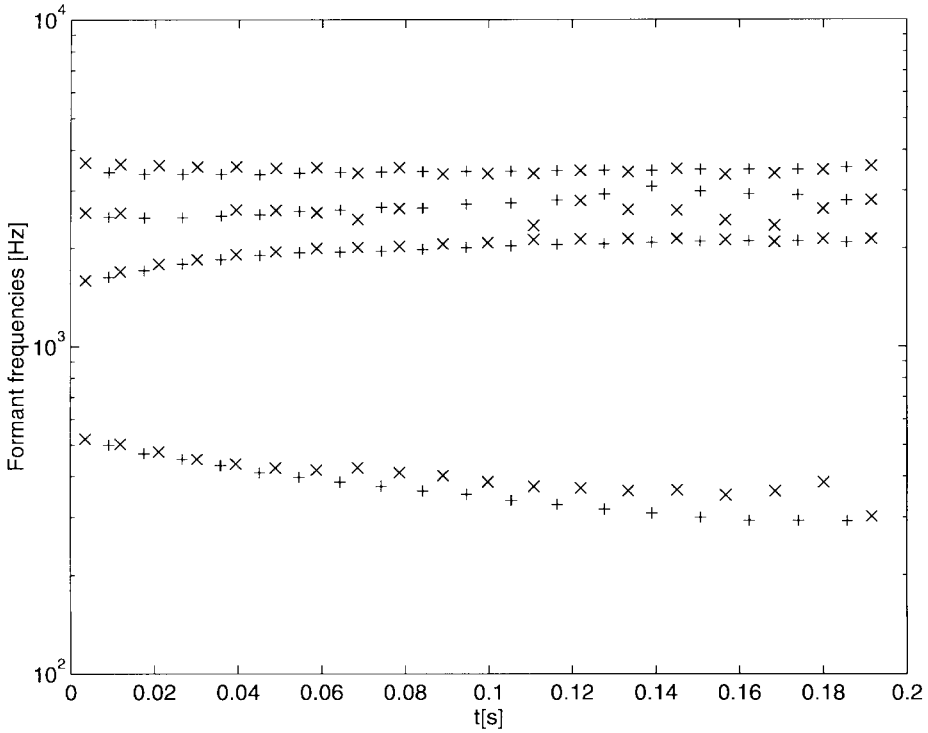

Fig. 4. Formant frequency tracks of male diphthong /eI/. Order of prediction: 9. Frame length: 5 ms. Symbols "×": preexcitation formant frequencies, estimated from the first-order difference signal. Frame ending 1.25 ms before the instant of excitation. Symbols "+": postexcitation formant frequencies.

Next, we include preexcitation formant frequencies. In order to avoid the influence of rapid changes in the glottal pulse, the preexcitation analysis frame ended an arbitrary 1.25 ms before the instant of excitation. Straightforward covariance analysis on these samples may not yield good estimates for two reasons. One is that, due to the higher damping of the resonances in the open phase, the formant amplitudes are usually much smaller than those in the closed phase. Another reason is the influence of the glottal pulse in the speech signal, which can be reduced by a preemphasis filter. Fig. 4 shows the results obtained with 5-ms analysis frames. The plot symbols "+" show the postexcitation formant frequencies. The plot symbols "×" show the preexcitation formant frequencies. The differences are mainly in the first formant, which is lower in the postexcitation case. This observation is in accordance with the results presented in [14]. The third formant is often missed in the preexcitation measurement, which is due to its higher bandwidth.

The necessity of the use of the multicycle covariance method in the case of high-pitched voices is illustrated by Fig. 5, which shows the waveform of a vowel /u/ uttered by a female (top panel), the postexcitation formant tracks obtained with the covariance method (middle panel), and the postexcitation formant tracks obtained with the multicycle covariance method (bottom panel). The sampling frequency of the signal was 8 kHz. In both cases, the analysis frames were 2.5 ms long. The multicycle covariance analysis was performed on three analysis frames. The formant tracks obtained with the covariance method show a lot of irregularities and formants are often missed. The reason for the irregularities in the formant tracks is probably the noisy glottal waveform. This is confirmed by the average normalized correlation factor between analysis frames, which was estimated as 0.65. For another /u/, recorded under the same conditions and uttered by a male speaker, this factor was 0.88. The formant tracks obtained with the multicycle covariance method are more consistent and fewer formants are missed.

When the postexcitation formant bandwidths were measured from the zeros of the prediction polynomial, we found them to be rather irregular as functions of time. This was even more so for the preexcitation formant bandwidths. We therefore restricted ourselves to postexcitation formant bandwidths and used the multicycle covariance method with three analysis frames, which has a smoothing effect. Fig. 6 shows the bandwidths of the first four postexcitation formants of the diphthong /eI/, obtained with a 3.25-ms analysis frame. All bandwidths are fairly smooth and consistent, though not as consistent as the measured postexcitation formant frequencies. The measured bandwidths of the first three formants confirm the observations regarding formant bandwidth made in [6]. The bandwidth of the fourth formant seems too small. A possible explanation for this is that this formant is close to a harmonic of the fundamental frequency. Anyway, it has been observed earlier that the extraction of formant bandwidths from prediction polynomials may be unreliable [12].

B. Comparison with Regularly Spaced Analysis Frames

Fig. 7 shows formant tracks obtained with 2.5-, 5-, and 10-ms analysis frames, respectively. The plot symbols "·" show formant frequencies measured from uniformly spaced,

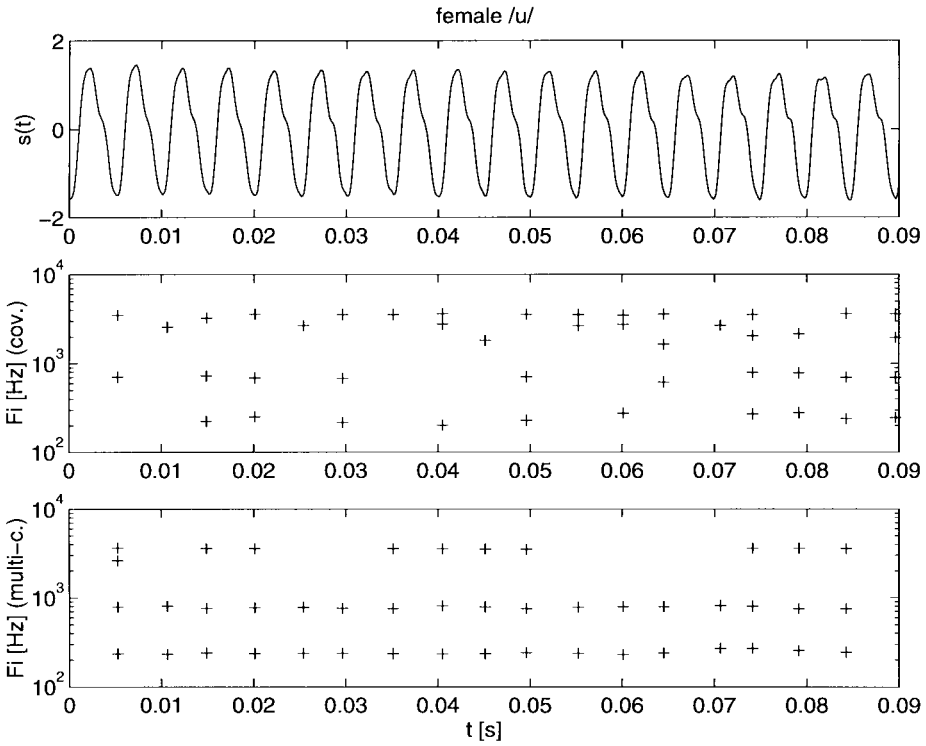

Fig. 5. Female vowel /u/. Top: waveform $s(t)$; sampling frequency, 8 kHz. Middle: postexcitation formant frequency tracks F_i, obtained with the covariance method; order of prediction, 9; frame length, 2.5 ms. Bottom: postexcitation formant frequency tracks F_i, obtained with the multicycle covariance method; number of analysis frames included, 3; order of prediction, 9; frame length, 2.5 ms.

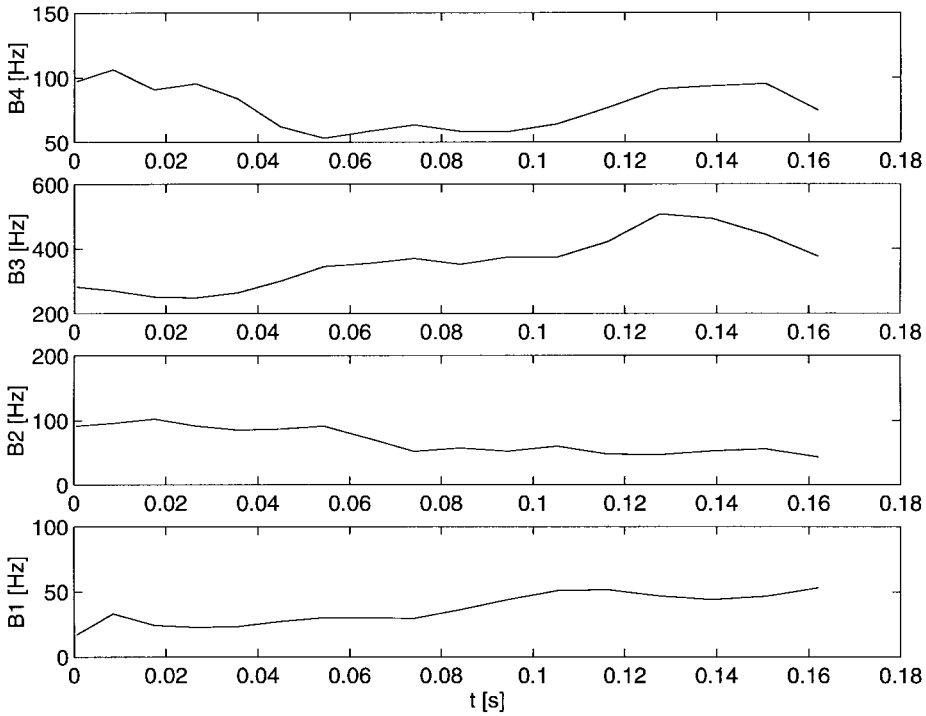

Fig. 6. Postexcitation formant bandwidth tracks B_1–B_4 of male diphthong /eI/, obtained with the multicycle covariance method. Number of analysis frames included: 3. Order of prediction: 9. Frame length: 3.75 ms.

consecutive analysis frames and the plot symbols "+" show postexcitation formant frequencies. For the shorter analysis frames of 2.5 and 5.0 ms, we observe that the postexcitation formant frequencies show a better consistency. For the 10-ms frames, the formant frequencies obtained with uniformly spaced frames are nearly as consistent as postexcitation formant frequencies, but the third formant is often missed due to its higher bandwidth.

C. Noise Sensitivity

The extraction of the instants of significant excitation is based on the assumption that the speech signal contains

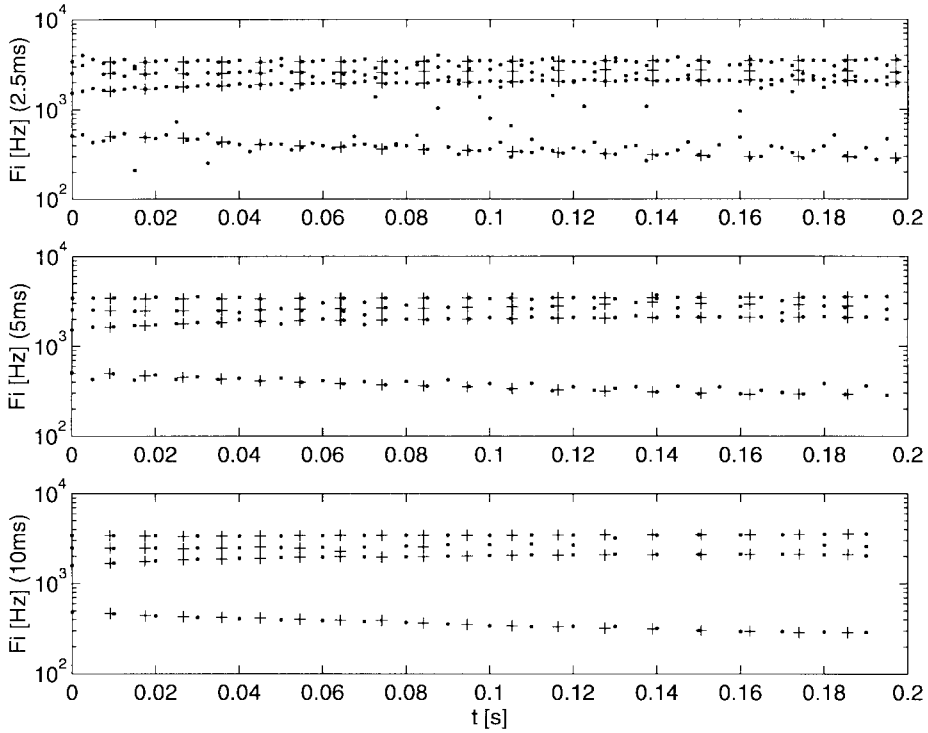

Fig. 7. Formant frequency tracks F_i of male diphthong /eI/ for analysis frame lengths of 2.5 ms (top), 5 ms (middle), and 10 ms (bottom). Order of prediction: 9. Symbols "·": uniformly spaced consecutive analysis frames. Symbols "+": postexcitation analysis frames.

delayed versions of minimum-phase impulse responses. There are a number of situations where this assumption may be violated, e.g., distortions due to nonminimum-phase filtering in the recording system or additive noise. This may lead to errors in the estimates of the instants of significant excitation. Formant parameter estimation with short postexcitation analysis frames requires reliable estimates of these instants. A small error of a few sample periods can have the effect that the analysis frame includes the moment of excitation, causing the minimization of (11) to produce an erroneous result. Delaying the analysis frames by a few samples reduces this sensitivity. We will briefly discuss the influence of additive white Gaussian noise on the extraction of the instants of significant excitation and on the estimation of the formant frequencies. An extensive discussion on the noise sensitivity of the estimation of the instants of significant excitation will be given in [27].

We first examined the errors in the instants of excitation due to pseudorandom additive white Gaussian noise at signal-to-noise ratios (SNR's) of 20–90 dB on a set of six male and six female utterances of the vowels /a/, /e/, /i/, /o/, /u/, and /y/, recorded at a sampling frequency of 8 kHz. The male pitch period was about 9 ms, and the female pitch period about 5 ms. Three types of errors occurred: i) erroneous instants of significant excitations were introduced; ii) instants of significant excitation remained undetected; and iii) the position of the estimated instant was shifted by a few samples. Errors of type i) will lead to wrongly placed analysis frames and are therefore unacceptable. Occasional errors of type ii) will lead to gaps in formant tracks that can be repaired with a formant-smoothing algorithm [12]. Errors of type iii) are acceptable if they are smaller than the number of samples by which the analysis frame has been delayed. Errors of types i) and ii) only occurred for SNR's below 40 dB. Above this value, we found errors in the estimated instants of excitation of at most two sample periods. This means that, when the present method is applied to noisy speech, the SNR should be better than 40 dB and a delay of at least two samples of the postexcitation analysis frame is required.

Some thoughts on the influence of additive noise on the estimation of the roots of prediction polynomials and further references are given in [28]. Here, we will only consider the effect of additional white Gaussian noise at an SNR of 40 dB on the estimates of formant frequencies. The value of 40 dB corresponds to the just-acceptable performance level of the instant estimation. The noise was added to 800-ms realizations of the six male and six female vowels /a/, /e/, /i/, /o/, /u/, and /y/. For each realization the root-mean-squared errors in the first and second formant with respect to the noiseless case were computed. The calculations were done using single postexcitation analysis frames, resulting in numbers $\sigma_1(F_1)$ and $\sigma_1(F_2)$, and using the multicycle covariance method, with three postexcitation analysis frames, resulting in numbers $\sigma_3(F_1)$ and $\sigma_3(F_2)$. In all cases, the analysis frames were 2.5 ms (20 samples) long. The results of 20 such trails were averaged. Results for the higher formants were not computed, since these formants were occasionally missed. The results are given in Table I.

The effect of the noise is greater on the estimated formants of female voices than on those of male voices. With four exceptions, it is also greater for the second than for the first formant. Table I illustrates the usefulness of the multicycle covariance method for the estimation of formant parameters in

TABLE I
AVERAGE ROOT-MEAN-SQUARED ERRORS $\overline{\sigma_n(F_k)}$ IN THE FIRST TWO FORMANTS OF MALE AND FEMALE VOWELS DUE TO ADDITIVE GAUSSIAN WHITE NOISE AT AN SNR OF 40 dB

		single cycle		multi cycle	
voice	vowel	$\sigma_1(F_1)$ [Hz]	$\sigma_1(F_2)$ [Hz]	$\sigma_3(F_1)$ [Hz]	$\sigma_3(F_2)$ [Hz]
male	/a/	3.2795	3.4031	0.2907	0.4496
	/e/	2.0617	10.1979	0.3346	1.0187
	/i/	8.2380	6.4577	0.2092	1.9267
	/o/	10.7935	17.0217	3.3521	5.3768
	/u/	8.0057	33.0739	5.2129	8.2553
	/y/	6.5897	15.0058	0.3337	2.3539
female	/a/	7.9829	9.5497	0.7410	1.7955
	/e/	23.0070	7.8000	5.4919	2.8430
	/i/	26.2573	49.9237	0.4193	2.0706
	/o/	36.4544	13.7153	2.1421	23.2044
	/u/	29.0033	59.0676	9.3912	14.6525
	/y/	19.7591	32.1282	0.6257	11.6487

noisy speech, which reduces the error by sometimes an order of a magnitude. The only exception is the second formant of the female vowel /o/. Inspection of the formant tracks of this utterance showed that the second formant was often missed and the third formant was used instead, which lead to the increased error. With the multicycle covariance method an error of only a few Hz can be achieved for the first and the second formant. Only in some cases the error in the second formant will be higher up to about 25 Hz. When the SNR was increased to 50 dB, all the errors in the multicycle case decreased by about a factor of three. In the single-cycle case the improvement was around a factor of two or sometimes less.

D. Discussion

The use of postexcitation analysis frames, which are shorter than one pitch period and preferably shorter than the closed phase, results in reliable, consistent estimates of the formant frequencies and to a lesser extent of the formant bandwidths. If the closed phase is short or the signal is expected to be noisy, the use of the multicycle covariance method is recommended. The multicycle covariance method has also been found to improve postexcitation formant-bandwidth estimation of stationary vowels. Pre- and postexcitation formant frequencies can attain different values. Preexcitation formant frequencies also seem to be less consistent. The use of short postexcitation analysis frames make it possible to capture high-bandwidth formants that cannot be revealed by methods based on larger analysis frames. The irregularity of the formant frequencies estimated with regularly spaced frames compared with postexcitation formant frequencies can be explained by the changing relative position of the frame in a pitch period in which formant frequencies fluctuate. Using larger frames will decrease this effect, because within the larger frame the closed-phase resonances with their lower damping will dominate. However, the use of larger frames may lead to the loss of formants with a higher damping.

V. RESULTS

In this section, we discuss the results of formant-frequency and bandwidth estimation for different types of speech signals, synchronizing the analysis frames with the instants of significant excitation. We will examine vowels, consonant-vowel transitions, and short fragments of sentences produced by both male and female speakers. The results are presented to demonstrate the possibilities of the analysis method presented in this paper and are not intended to be general.

A. Vowels

We used the male and female vowels /a/, /i/, and /u/ from the set that was described in Section IV.

Postexcitation analysis frames started immediately after the instants of excitation and preexcitation analysis frames ended 0.6125 ms before them. The preexcitation analysis frames were taken from the first-order difference signal. The male voice was analyzed with a frame size of 3.75 ms, which is well below the pitch period. The female voice was analyzed with a frame size of 2.5 ms and the multicycle covariance method with three analysis frames was applied. In a first measurement, the pre- and postexcitation formant frequencies of the male voice were often missed. This improved when we also applied the multicycle covariance method with three analysis frames to the male voice. The multicycle covariance method was also used for the formant-bandwidth estimation. Formant bandwidth estimation was only done for the vowel /a/, because too many formants were missed in the other cases.

The results of the formant frequency measurement are shown in Fig. 8, left panel, for the male voice, and in Fig. 8, right panel, for the female voice. The plot symbols "+" show postexcitation formant frequencies and the plot symbols "·" show preexcitation formant frequencies.

In all cases, the postexcitation formant frequencies are consistent in time. For the vowels /u/, occasionally the third formant is missed. Inspection shows that this is due to the formant bandwidth, which in those cases exceeds the limit of 570 Hz. The preexcitation formants are more irregular. Sometimes an entire formant track is missing. When the preexcitation-formant tracks are consistent, they show an upward shift of the lowest formant in the case of the male /a/, /i/, and /u/. This phenomenon is also present in the female /a/ and /i/. In the female /u/, the lowest formant could not be measured in this way.

The results of the formant bandwidth measurement are shown in Fig. 9, left panel, for the male voice, and in Fig. 9, right panel, for the female voice. The solid lines show the postexcitation formant bandwidths, and the dashed lines show the preexcitation formant bandwidths.

The postexcitation formant bandwidths of the male vowel /a/ seem fairly constant. The preexcitation formant bandwidths are slightly more irregular. Furthermore, the preexcitation formant bandwidths are always a little above the postexcitation formant bandwidths. The formant bandwidths of the female vowel /a/ are in all cases more irregular than those of the male vowel /a/. The preexcitation formant bandwidths of the female vowel /a/ are not systematically above the postexcitation formant bandwidths. The measured bandwidths are somewhat higher than those presented in [6] and [29]. In order to check whether this was due to a possible smearing effect of the multicycle covariance method, we also computed the

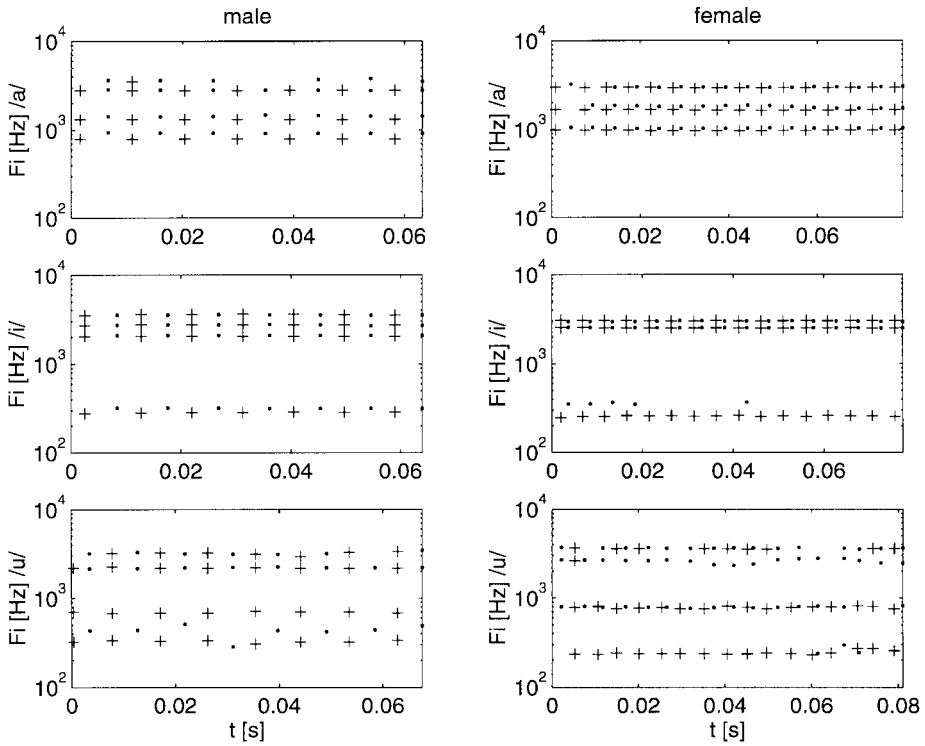

Fig. 8. Formant frequency tracks F_i of male and female vowels /a/, /i/, and /u/. Sampling frequency: 8 kHz; order of prediction, 9. Obtained with the multicycle covariance method. Number of combined pitch analysis frames: 3. Left: male vowels; frame length, 3.75 ms. Right: female vowels; frame length, 2.5 ms. Symbols "·": preexcitation formant frequencies estimated from the first-order difference signal, the analysis frame ending 0.6125 ms before the instant of excitation. Symbols "+z": postexcitation formant frequencies.

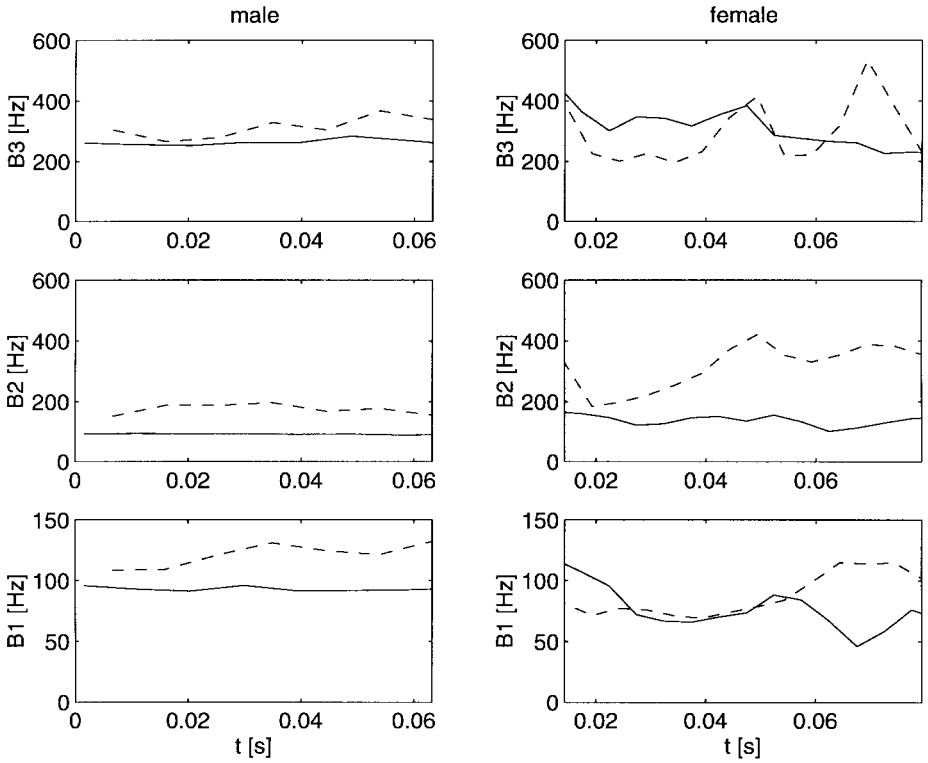

Fig. 9. Formant bandwidth tracks B_1–B_3 of male and female vowels /a/. Sampling frequency: 8 kHz. Order of prediction: 9. Obtained with the multicycle covariance method. Number of combined pitch analysis frames: 3. Left: male vowels; frame length, 3.75 ms. Right: female vowels; frame length, 2.5 ms. Dashed line: preexcitation formant bandwidths estimated from the first-order difference signal, the analysis frame ending 0.6125 ms before the instant of excitation. Solid line: Postexcitation formant bandwidths.

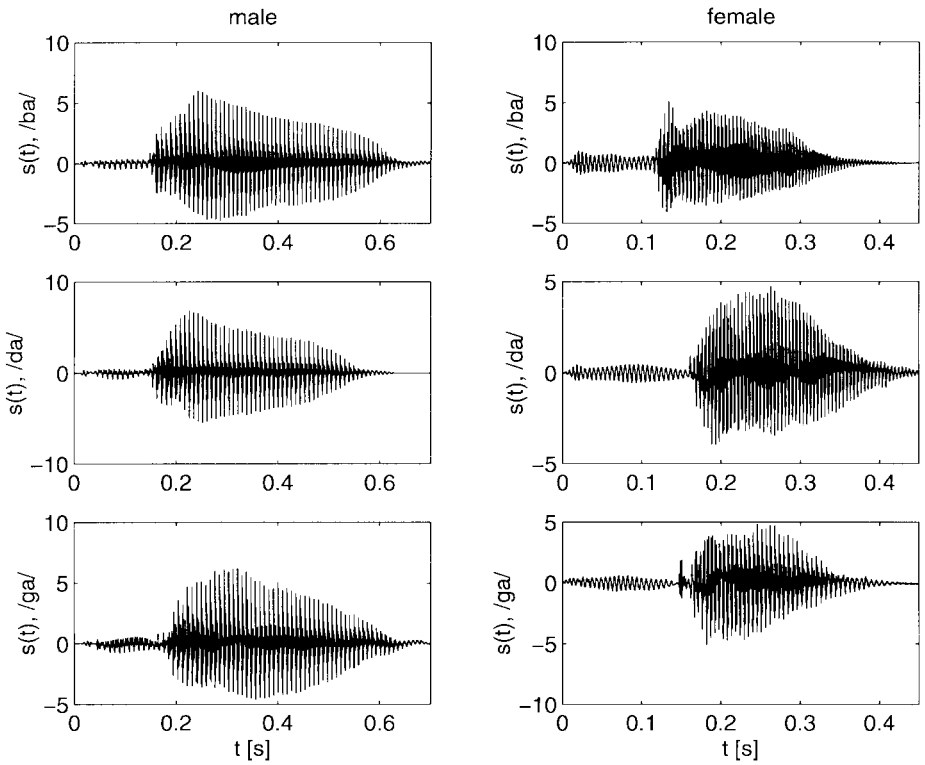

Fig. 10. Male and female consonant-vowel transitions /ba/, /da/, and /ga/. Sampling frequency: 8 kHz.

postexcitation formant bandwidths from single glottal cycles and from the entire 800 ms segments. All results appeared to be similar, although the single-cycle tracks were more irregular. The differences between the results presented here and those in [6] and [9] may be explained by the different measurements methods that have been used. In the present case, the formant resonances are obtained directly from the speech signal. The method described in [6] and [9] fits a resonance curve to an estimated spectrogram and subsequently computes formant frequencies and bandwidths from this curve.

B. Consonant-Vowel Transitions

Fig. 10 shows the time signals of the consonant-vowel transitions /ba/, /da/, and /ga/ that were used. The sampling frequency was 8 kHz. Because of the irregularities observed in the preexcitation formant frequencies and in pre- and postexcitation formant bandwidths, we have not included these measurements for the consonant-vowel transitions, but only consider postexcitation formant frequencies. The analysis parameters were the same as those used for the vowels. The multicycle covariance method with three analysis frames was used for the female voice. The results for the male voice are shown in Fig. 11, left panel. The results for the female voice are shown in Fig. 11, right panel.

The fourth formant of the male voice is sometimes missed. For the female voice it is always missed, and the third formant is occasionally missed as well. In some of the formants of the female voice, a minor flattening-off effect can be observed immediately after the release. This may be due to the multicycle covariance method, since the effect lasts for three cycles and is not present in the formants obtained from the male voice. The regularity of the formant transition and the consistency of the proposed method demonstrates the utility of the method in extracting the dynamic characteristics of the vocal-tract system for the most difficult segments of speech signals. Note the irregular formant locations that appear during and after the decay of the vowel /a/.

C. Fragments of Sentences

We only consider postexcitation formant frequencies. The analysis parameters were the same as those used for the vowels. The multicycle covariance method with three analysis frames was used for the female voice. Fig. 12 shows the time signal of the utterance “any dictionary,” spoken by a male voice, and the corresponding formant tracks. The sampling frequency was 10 kHz. Fig. 13 shows the time signal of the same utterance spoken by a female voice, and the corresponding formant tracks. The sampling frequency was also 10 kHz.

The formant tracks obtained from the fragments of sentences seem more scattered than the tracks obtained from the vowels or the consonant-vowel transitions. This is because the estimates corresponding to unvoiced or silent parts have not been removed, nor has any form of smoothing been applied.

VI. Conclusions

In this paper, we have demonstrated that an accurate analysis of speech signals is essential to extract the dynamic characteristics of the vocal-tract system. Such an analysis is possible by a proper choice of the size and positions of the analysis frames. We have shown that synchronizing analysis frames with the

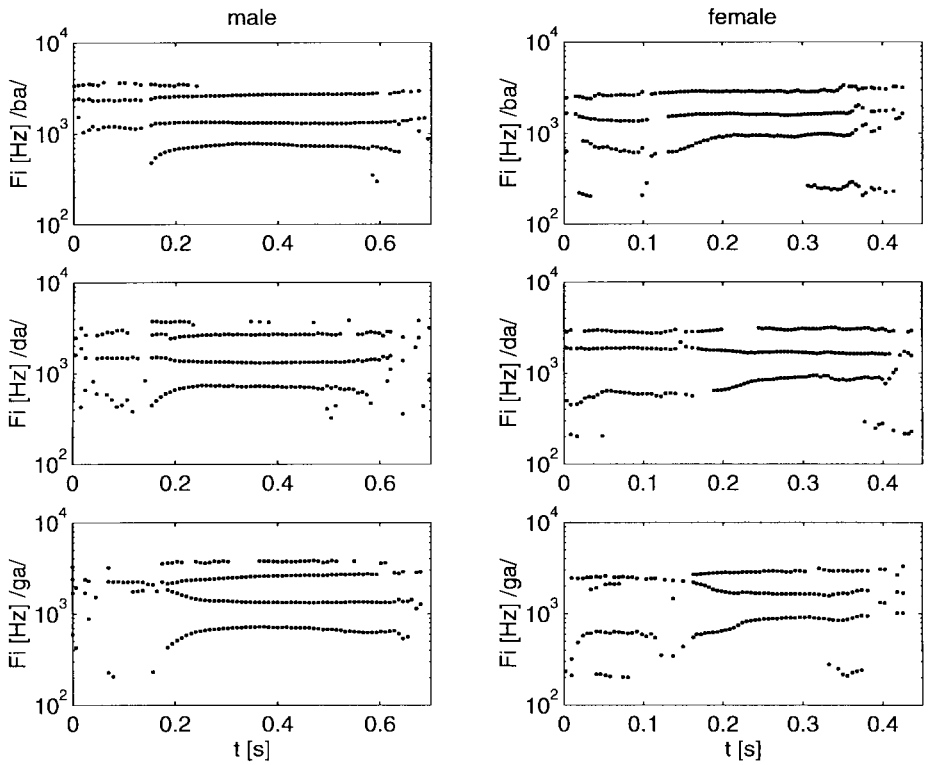

Fig. 11. Postexcitation formant frequency tracks F_i of male and female consonant-vowel transitions /ba/, /da/, and /ga/. Order of prediction: 9. Left: male consonant-vowel transitions; frame length, 3.75 ms. Right: female consonant-vowel transitions; frame length, 2.5 ms. Obtained with the multicycle covariance method. Number of combined pitch analysis frames: 3.

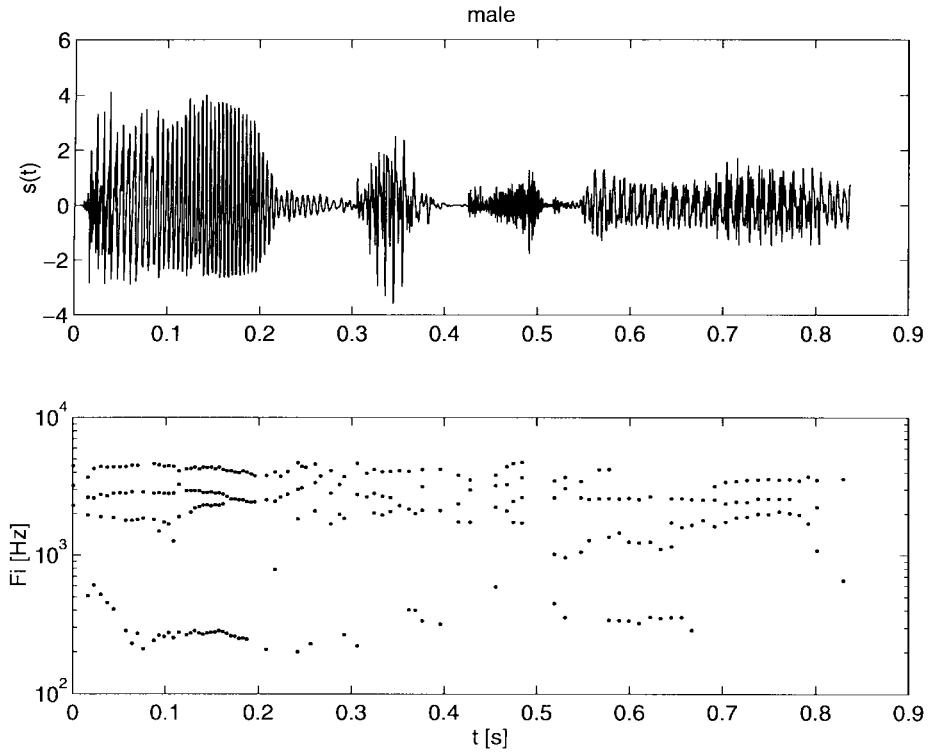

Fig. 12. Male utterance "any dictionary." Top: waveform $s(t)$; sampling frequency, 10 kHz. Bottom: postexcitation formant frequency track F_i; frame length, 3.75 ms; order of prediction, 9.

instants of glottal closure gives highly consistent estimates of the formant frequencies. To a lesser extent, this is also the case for the formant bandwidths. We have distinguished pre- and postexcitation measurements in which the analysis frames are positioned immediately before or after the instant of glottal closure. It can be desirable to choose the frame sizes such that the analysis frames fit in the open or closed glottal phases. When short analysis frames are required, reliable estimates

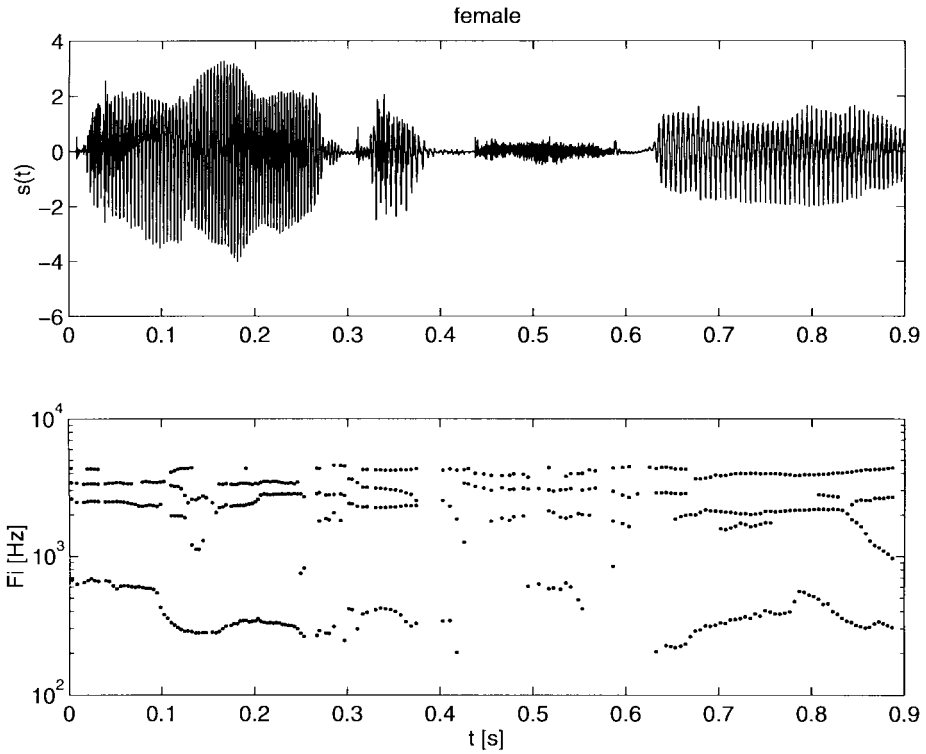

Fig. 13. Female utterance "any dictionary." Top: waveform $s(t)$; sampling frequency: 10 kHz. Bottom: postexcitation formant frequency track F_i; frame length, 2.5 ms; Order of prediction, 9. Obtained with the multicycle covariance method. Number of analysis frames included: 3.

for formant frequencies or bandwidths can still be obtained by applying the multicycle covariance method, which combines a number of analysis frames. This method can also be used to improve the estimates when the signal is noisy. A particular advantage of the method presented here over other methods is that short postexcitation analysis frames make it possible to track high-bandwidth formants.

Preexcitation formant-frequency tracks were less consistent than postexcitation formant-frequency tracks and more formants were missed. However, in many cases this analysis can still be performed. For certain vowels the pre- and postexcitation formant frequencies can differ significantly. The estimated formant-bandwidth tracks were found to be less consistent than estimated formant-frequency tracks, but for stationary vowels consistent estimates of formant bandwidths can still be obtained.

In general, the most consistent results were obtained from postexcitation analysis frames. A possible explanation is that the glottal airflow is smallest or zero immediately after glottal closure. This means that the effect of the nonlinear coupling between the sub- and the supraglottal tract and the influence of the subglottal tract on the formant frequencies are smaller in the postexcitation analysis frames than in the preexcitation frames.

REFERENCES

[1] V. N. Sorokin, "Determination of vocal tract shape for vowels," *Speech Commun.*, vol. 11, pp. 71–85, 1992.

[2] ——, "Inverse problems for fricatives," *Speech Commun.*, vol. 14, pp. 249–262, 1994.

[3] T. V. Ananthapadmanabha and G. Fant, "Calculations of true glottal volume-velocity and its components," *Speech Commun.*, vol. 1, pp. 167–184, 1982.

[4] B. Cranen and L. Boves, "On subglottal formant analysis," *J. Acoust. Soc. Amer.*, vol. 81, pp. 734–746, 1987.

[5] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 663–671, July 1994.

[6] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. New York: Springer-Verlag, 1965.

[7] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1970.

[8] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, pp. 1233–1268, 1972.

[9] R. N. J. Veldhuis, I. J. M. Bogaert, and N. J. C. Lous, "Two-mass models for speech synthesis," in *Proc. Eurospeech'95*, Madrid, Spain, pp. 1853–1856.

[10] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 350–355, Aug. 1979.

[11] D. H. Klatt and L. C. Klatt, "Analysis synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, pp. 820–856, 1990.

[12] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

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

[14] G. Fant, K. Ishizaka, J Lindqvist, and J. Sundberg, "Speech analysis and speech production," Speech Transmission Lab. Quart. Progr. Rep. 1/72, KTH, 1972.

[15] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[16] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[17] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Commun.*, vol. 10, pp. 209–221, Aug. 1991.

[18] A. K. Krishnamurty and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 730–743, 1986.

[19] D. G. Childers and C. K. Lee, "Voice quality factors: Analysis synthesis and perception," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2394–2410, 1991.
[20] R. Smits and B. Yegnanayarana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 325–333, Sept. 1995.
[21] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. ICASSP-95*, Detroit, MI, 1995.
[22] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
[23] G. Fant, J. Liljencrants, and Q. Lin, "A four-paremeter model of glottal flow," Speech Transmission Lab. Quart. Progr. Rep. 4/85, KTH, 1985.
[24] R. Smits, "Accuracy of quasistationary analysis of highly dynamic speech signals," *J. Acoust. Soc. Amer.*, vol. 96, pp. 3401–3415, 1994.
[25] J. D. Markel, "Digital inverse filtering—A new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, pp. 129–137, 1972.
[26] D. O'Shaughnessy, *Speech Communication*. Reading, MA: Addison-Wesley, 1990.
[27] P. S. Murthy and B. Yegnanarayana, "Robustness of group delay based method for extraction of significant instants from speech signals," submitted for publication.
[28] S. L. Marple, Jr., *Digital Spectral Analysis with Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[29] H. K. Dunn, "Methods of measuring vowel formant bandwidths," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1737–1746, 1961.

B. Yegnanarayana (M'78–SM'84) was born in India on January 9, 1944. He received the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1964, 1966, and 1974, respectively.

He was a Lecturer from 1966 to 1974 and an Assistant Professor from 1974 to 1978 in the Department of Electrical Communication Engineering, Indian Institute of Science. From 1977 to 1980, he was a Visiting Associate Professor of Computer Science at Carnegie Mellon University, Pittsburgh, PA. He was a visiting scientist at ISRO Satellite Center, Bangalore, from July to December 1980. Since 1980, he has been a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology, Madras. He was a visiting Professor at the Institute of Perception Research, Eindhoven Technical University, Eindhoven, The Netherlands, from July 1994 to January 1995. From 1966 to 1971, he was engaged in the development of environmental test facilities for the Acoustics Laboratory, Indian Institute of Science. Since 1972, he has been working on problems in the area of speech signal processing. He is presently engaged in research activities in digital signal processing, speech recognition, and neural networks.

Dr. Yegnanarayana is a member of the Computer Society of India, a fellow of the Institution of Electronics and Telecommunication Engineers of India, and a Fellow of the Indian National Academy of Engineers.

Raymond N. J. Veldhuis was born in The Hague, The Netherlands, on April 8, 1955. He received the Engineer degree in electronics from the Twente University, The Netherlands, in 1981, and the Ph.D. degree from Nijmegen University, Nijmegen, The Netherlands, in 1988.

He joined Philips Research Laboratories, Eindhoven, The Netherlands, in 1981, where he has worked on audio signal restoration and audio source coding. In 1992 he joined the IPO-Research Center on User-System Interaction. At present he is engaged in speech signal processing and speech synthesis.