Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals

B. Yegnanarayana, *Senior Member, IEEE*, and K. Sri Rama Murty

***Abstract*—Exploiting the impulse-like nature of excitation in the sequence of glottal cycles, a method is proposed to derive the instantaneous fundamental frequency from speech signals. The method involves passing the speech signal through two ideal resonators located at zero frequency. A filtered signal is derived from the output of the resonators by subtracting the local mean computed over an interval corresponding to the average pitch period. The positive zero crossings in the filtered signal correspond to the locations of the strong impulses in each glottal cycle. Then the instantaneous fundamental frequency is obtained by taking the reciprocal of the interval between successive positive zero crossings. Due to filtering by zero-frequency resonator, the effects of noise and vocal-tract variations are practically eliminated. For the same reason, the method is also robust to degradation in speech due to additive noise. The accuracy of the fundamental frequency estimation by the proposed method is comparable or even better than many existing methods. Moreover, the proposed method is also robust against rapid variation of the pitch period or vocal-tract changes. The method works well even when the glottal cycles are not periodic or when the speech signals are not correlated in successive glottal cycles.**

***Index Terms*—Autocorrelation, fundamental frequency, glottal closure instant, periodicity, pitch, zero-frequency resonator.**

I. Introduction

VOICED sounds are produced from the time-varying vocal-tract system excited by a sequence of events caused by vocal fold vibrations. The vibrations of the vocal folds result in a sequence of glottal pulses with major excitation taking place around the instant of glottal closure (GCI). The rate of vibration of the vocal folds determines the fundamental frequency (F_0), and contributes to the perceived pitch of the sound produced by the vocal-tract system. Though the usage of the term "rate of vibration" gives an impression that the vibrations of the vocal folds are periodic, in practice the vocal fold vibrations at the glottis may or may not be periodic. Even a periodic vibration of the vocal folds at the glottis may produce a speech signal that is less periodic because of the time-varying vocal-tract system that filters the glottal pulses. Sometimes, the vocal fold vibrations at the glottis themselves may show aperiodic behavior, as in the case of changes in the shape of the glottal flow waveform (for example, the changes in the duty cycles of open/closed phases), or the intervals where the vocal fold vibration reflect several superposed periodicities (diplophony) [1], or where the glottal pulses occur without obvious regularity in the time (glottalization, vocal fry, or creaky voice) [2]. In practice, the rate of vibration of the vocal folds may change from one glottal cycle to the next cycle. Hence, it is more appropriate to define the instantaneous fundamental frequency of excitation source for every glottal cycle. In this paper, we propose an event-based approach to accurately estimate the instantaneous fundamental frequency from speech signals. Throughout the paper, we use the terms fundamental frequency and pitch frequency interchangeably.

Accurate estimation of the fundamental frequency of voiced speech plays an important role in speech analysis and processing applications. The variation in the fundamental frequency with time contributes to the speech prosody. Estimation of accurate prosody is useful in various applications such as in speaker recognition [3], [4], language identification [5], and even speech recognition [6], [7]. Prosody also reflects the emotion characteristics of a speaker [8]. Prosody is essential for producing high-quality speech synthesis, and also for voice conversion. Prosody features were exploited for hypothesizing sentence boundaries [9], for speech segmentation, and for story parsing [10]. Although many methods of pitch estimation have been proposed, reliable and accurate detection is still a challenging task, especially when the speech signal is weakly periodic, and the instantaneous values of pitch vary even within an analysis frame consisting of a few glottal cycles. The presence of noise in the speech signal further complicates the problem of pitch estimation, and degrades the performance of the pitch estimation algorithms.

There are several algorithms proposed in the literature for estimating the fundamental frequency from speech signals [11]–[13]. Depending on the type of processing involved, the algorithms may be classified into three broad categories: 1) algorithms using time domain properties; 2) algorithms using frequency domain properties; and 3) algorithms using statistical methods to aid in the decision making [14]–[16].

Algorithms based on the properties in the time domain operate directly on the speech signal to estimate the fundamental frequency. Depending on the size of the segment used for processing, the time domain methods can be further categorized into *block-based* methods and *event-based* methods. In the block-based methods, an estimate of the fundamental frequency is obtained for each segment of speech, where it is assumed that the pitch is constant over the segment consisting

Manuscript received June 03, 2008; revised November 25, 2008. Current version published March 18, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

B. Yegnanarayana is with International Institute of Information Technology, Hyderabad 500 032, India (e-mail: yegna@iiit.ac.in).

K. Sri Rama Murty is with the Department of Computer Science and Engineering, Indian Institute of Technology-Madras, Chennai 600 036, India (e-mail: ksrmurty@gmail.com).

Digital Object Identifier 10.1109/TASL.2008.2012194

of several pitch periods. In this case, variations of the fundamental frequency within the segment are not captured. Among the time domain block-based methods, the autocorrelation approaches are popular for their simplicity. For a periodic signal, its autocorrelation function is also periodic. Due to periodic nature of the voiced speech, the first peak (also called the pitch peak) after the center peak in the autocorrelation function indicates the fundamental period (T_0) of the signal. The reciprocal $F_0 = (1/T_0)$ is the fundamental frequency. The main limitation of this method is that the pitch peak may get obscured due to the presence of other spurious peaks. The spurious peaks may arise due to noise, or due to the formant structure of the vocal-tract system, or due to the quasi-periodic nature of the speech signal, or due to the position and length of the analysis window.

Event-based pitch detectors estimate the pitch period by locating the instants at which glottis closes (called events), and then measuring the time interval between two such events. Wavelet transforms are used for pitch period estimation based on the assumption that the glottal closure causes sharp discontinuities in the derivative of the airflow [17]. The transients in the speech signal due to glottal closure result in maxima in the scales of the wavelet transform around the instant of discontinuity. An optimization scheme is proposed in the wavelet framework using a multipulse excitation model for speech signal, and the pitch period is estimated as a result of this optimization [18]. An instantaneous pitch estimation algorithm which exploits the advantages of both block-based and event-based approaches is given in [19]. In this method, the pitch is modeled by a B-spline expansion, which is optimized using a multistage procedure for improving the robustness.

Algorithms based on the properties in the frequency domain assume that if the signal is periodic in the time domain, then the frequency spectrum of the signal contains a sequence of impulses at the fundamental frequency and its harmonics [20]–[23]. Then simple measurements can be made on the frequency spectrum of the signal, or on a nonlinearly transformed version of it, to estimate the fundamental frequency of the signal. The cepstrum method for extraction of pitch utilizes the frequency domain properties of speech signals [20]. In the short-time spectrum of a given voiced frame, the information about the vocal-tract system appears as a slowly varying component, and the information of the excitation source is in rapidly varying component. These two components may be separated by considering the logarithm of the spectrum, and then applying the inverse Fourier transform to obtain the cepstrum. This operation transforms the information in the frequency domain to the cepstral domain, which has a strong peak at the average fundamental period of the voiced speech segment being analyzed.

Simplified inverse filter tracking (SIFT) algorithm uses both time and frequency domain properties of the speech signal [24]. In the SIFT algorithm, the speech signal is spectrally flattened approximately, and autocorrelation analysis is used on the spectrally flattened signal to extract pitch. Due to spectral flattening, a prominent peak will be present in the autocorrelation function at the pitch period of the voiced speech frame being analyzed.

Most of the existing methods for extraction of the fundamental frequency assume periodicity in successive glottal cycles, and therefore they work well for clean speech. The performance of these methods is severely affected if the speech signal is degraded due to noise or other distortions. This is because the pitch peak in the autocorrelation function or cepstrum may not be prominent or unambiguous. In fact, during the production of voiced speech, the vocal-tract system is excited by a sequence of impulse-like signals caused by the rapid closure of the glottis in each cycle. There is no guarantee that the physical system, especially due to the time-varying vocal-tract shape, produces similar speech signal waveforms for each excitation. Moreover, there is also no guarantee that the impulses occur in the sequence with any strict periodicity. In view of this, it is better to extract the interval between successive impulses, and take the reciprocal of that interval as the instantaneous fundamental frequency. In the next section, the basis for the proposed method of fundamental frequency estimation is discussed. In Section III, a method for pitch extraction from the speech signals is developed. In Section IV, the proposed method is compared with some standard methods for pitch extraction on standard databases, for which the ground truth is available in the form of electroglottograph (EGG) waveforms. The performance of the proposed method is also evaluated for different cases of simulated degradations in speech. Finally, in Section V, a summary of the ideas presented in this paper is given along with some issues that need to be addressed while dealing with speech signals in practical environments.

II. Basis for the Proposed Method of Pitch Estimation

As mentioned earlier, voiced speech is the output of the time-varying vocal-tract filter excited by a sequence of glottal pulses caused by the vocal fold vibrations. The vocal-tract system modulates the excitation source by formant frequencies, which depend on the sound unit being generated. The formant frequencies together with the fundamental frequency form important features of the voiced speech. There is an important distinction in the production of a formant frequency and in the production of the fundamental frequency. Formant frequencies are due to resonances of the vocal-tract system. The frequency of the resulting damped sinusoids are controlled by the size and the shape of the vocal-tract through the movement of the articulators. Because of the damped sinusoidal nature of the resonance, the formant frequency appears as a broad resonant peak in the frequency domain, but the fundamental frequency or pitch is perceived as a result of vibration of the vocal folds, which produces a sequence of regularly spaced impulses over short intervals of time. Periodic sequence of impulses in the time domain results in a periodic sequence of impulses in the frequency domain also. Hence, unlike the formant frequency, the information about the fundamental frequency is spread across the frequency range. This redundancy of information about the fundamental frequency in the frequency domain makes it a robust feature for speech analysis. For example, this redundancy helps us in perceiving the pitch even when the fundamental frequency is not present in the speech signal (as in the case of telephone speech).

It appears that in speech production mechanism the energy in the higher (>300 Hz) frequencies is produced in the form of formants, whereas the perception of low (<300 Hz) frequencies is primarily due to the sequence of glottal cycles. In fact, the perception of pitch (<300 Hz) is felt more due to the intervals between the impulses rather than due to presence of any low-frequency components in the form of sinusoids. In other words, it is the strong discontinuities at these impulse locations in the sequence that are producing the low-frequency effect in perception. Moreover, the information about the discontinuities is spread across all the frequencies including the zero frequency. In this paper, we propose a method based on using a resonator located at the zero frequency to derive the information about the impulse-like discontinuity in each glottal cycle. The derived sequence of impulse locations is used for estimating the fundamental frequency for each glottal cycle. Note that since the proposed method is based mainly on the assumption of a sequence (not necessarily periodic) of impulse-like excitation of the vocal-tract, it is better to interpret the operations in the time-domain. The frequency domain interpretation is not very relevant, and hence is used minimally throughout the paper. Moreover, due to dependence of the method on the impulse-like excitation, any spurious impulses caused by echoes or reverberation, or due to communication channels like telephone may affect the performance of the method. Also, in the case of telephone channels, the frequency components below 300 Hz are heavily damped (i.e., practically eliminated). The output of the zero-frequency filter may not bring out the effects due to impulse excitation. Hence, the proposed method may not work well for telephone and high-pass filtered speech signals.

III. Method for Estimating Fundamental Frequency From Speech Signals

A. Output of Zero-Frequency Resonator

The discontinuity due to an impulse excitation is reflected across all the frequencies including the zero frequency. That is, even the output of the resonator located at zero frequency should have the information of the discontinuities due to impulse-like excitation. We prefer to use the term *resonator*, even though ideally its location at zero frequency does not correspond to the normal concept of resonance. The advantage of choosing a filter at zero frequency is that the output of the resonator is not affected by the time-varying vocal-tract system. This is because the resonances of the vocal-tract system are located at much higher frequencies than at the zero frequency. Thus, the sequence of the excitation impulses, especially their locations, can be extracted by passing the speech signal though a zero-frequency filter. The signal is passed twice through the (zero-frequency) resonator to reduce the effects of all the resonances of the vocal-tract system. A cascade of two zero-frequency resonators provides a sharper cut-off compared to a single zero-frequency resonator. This will produce approximately a 24 dB per octave roll-off, thus damping out heavily all the frequency components beyond the zero frequency. Since the output of a zero-frequency resonator is equivalent to double integration of the signal, passing the speech signal twice though the zero-frequency resonator is equivalent to successive integration of the signal four times. This will result in a filtered output that has approximately polynomial growth/decay with time, as shown in Fig. 1(b). The effect of discontinuities due to impulse sequences will be overriding with small amplitude fluctuations on those large values of the output signal. We attempt to compute the deviation of the output signal from the local mean to extract the characteristics of the discontinuities due to impulse excitation. Note that the computation of the local mean is difficult due to rapid growth/decay of the output signal. This is the reason why it is preferable not to choose more than two resonators. The choice of the window length for computing the local mean depends on the interval between the discontinuities. A window length of about the average pitch period is used to compute the local mean. The resulting mean subtracted signal is shown in Fig. 1(c) for the speech signal in Fig. 1(a). We call the mean subtracted signal the "zero-frequency filtered signal" or merely the "filtered signal." The following steps are involved in processing the speech signal to derive the zero-frequency filtered signal [25], [26].

1) Difference the speech signal $s[n]$ to remove any dc or low-frequency bias during recording

$$x[n] = s[n] - s[n-1]. \tag{1}$$

2) Pass the differenced speech signal $x[n]$ twice through an ideal resonator at zero frequency. That is

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{2a}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \tag{2b}$$

where $a_1 = -2$, and $a_2 = 1$. Though, this operation is similar to successive integration of $x[n]$ four times, we prefer to interpret it as filtering at zero frequency.

3) Remove the trend in $y_2[n]$ by subtracting the mean over about 10 ms window at each sample. The resulting signal is given by

$$y[n] = y_2[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_2[n+m] \tag{3}$$

where $2N+1$ corresponds to the length of window in number of samples used to compute the mean. The resulting signal is called the filtered signal.

The filtered signal clearly shows rapid changes around the positive zero crossings. The locations of the instants of positive zero crossings in Fig. 1(c) are shown in Fig. 1(a) and (d) for comparison with discontinuities in the speech signal and in the differenced EGG waveform, respectively. There is close agreement between the locations of the strong negative peaks in the differenced EGG signal and the instants of positive zero crossings derived from the filtered signal. Therefore, the time instants of the positive zero crossings can be used as anchor points to estimate the fundamental frequency. The instants of positive zero

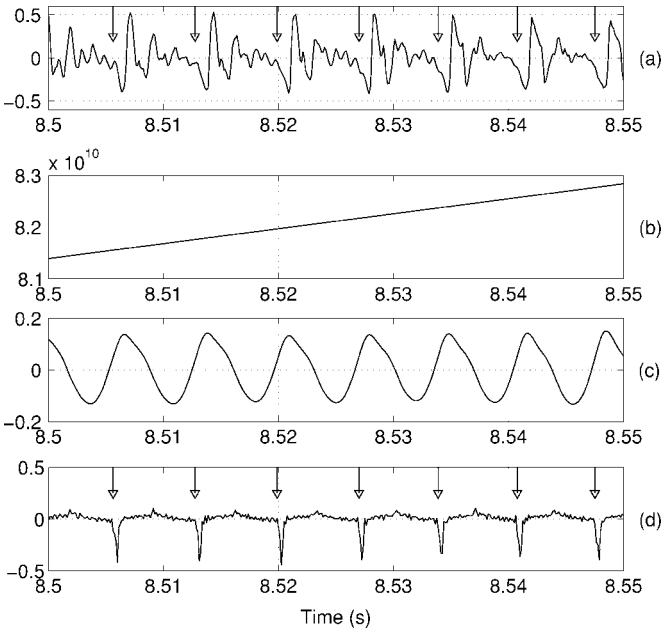

Fig. 1. (a) 50 ms segment of speech signal taken from continuous speech utterance. (b) Output of cascade of two zero-frequency resonators. (c) Filtered signal obtained after mean subtraction. (d) Differenced EGG signal. The locations of positive zero crossings in the filtered signal (c) are also shown in (a) and (d) for comparison with speech signal and differenced EGG signal.

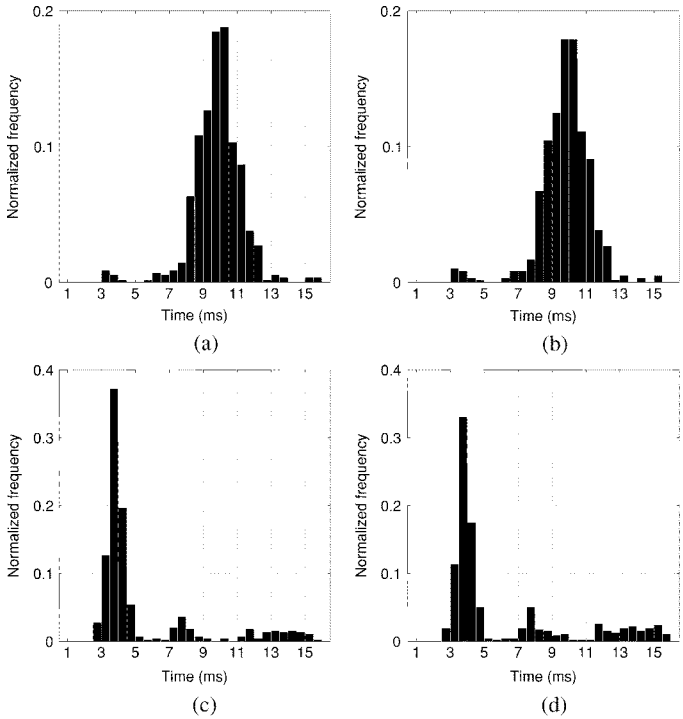

Fig. 2. Histogram of the locations of the pitch peak in the autocorrelation function for (a) clean signal from a male speaker, (b) speech signal from the same male speaker at 0 dB, (c) clean speech signal from a female speaker, and (d) speech signal from the same female speaker at 0 dB. Note that the location of the peak in the histogram plot is not affected by noise. (a) Male Speaker (Clean). (b) Male Speaker (0 dB). (c) Female Speaker (Clean). (d) Female Speaker (0 dB).

crossing of the filtered signal correspond to locations of the excitation impulses even when the impulse sequence is not periodic. It is important to note that such relation between the excitation signal and the filtered signal does not exist for a random noise excitation of the vocal-tract system. Also, the filtered signal will have significantly lower values for the random noise excitation compared to the sequence of impulse-like excitation. This is due to concentration of energy at the location of the impulse relative to the neighboring values. In the case of random noise there is no isolated impulse-like characteristic in the excitation.

B. Selection of Window Length for Mean Subtraction

To remove the trend in the output of the zero-frequency resonator, suitable window length need to be chosen to compute the local mean. The length of the window depends on the growth/decay of the output, and also on the overriding fluctuations in the output. The growth/decay in turn depends on the nature of the signal. The desired information of the overriding fluctuations depends on the intervals between impulses. If the window length is too small relative to the average duration (pitch period) between impulses, then spurious zero crossings may occur in the filtered signal, affecting the locations of the genuine zero crossings. If the window length is too large relative to the average pitch period, then also the genuine zero crossings are affected in the filtered signal. The choice of the window length for computing the local mean is not very critical, as long as it is in the range of about 1 to 2 times the average pitch period.

The average pitch period information can be derived in several ways. One way is to use the autocorrelation function of short (30 ms) segments of differenced speech, and determine the pitch period from the locations of the strongest peak in the interval 2 ms to 15 ms (normal range of pitch period). The histogram of the pitch periods is plotted. The pitch period value corresponding to the peak in the histogram can be chosen as the window length. Much simpler procedures can also be used to obtain an estimate of the average pitch period.

The average pitch period can be estimated using the histogram method even from degraded speech as shown in Fig. 2 for a male and a female speech at two different signal-to-noise ratios (SNRs). The location of the peak does not change significantly even under noisy conditions. Hence, the average pitch period can be estimated reliably. The filtered signal and the locations of the positive zero crossings in the filtered signal are shown in Fig. 3 for two different window lengths 7 ms and 16 ms for speech from a male voice having a pitch period of around 7 ms.

C. Validation of F_0 Estimates Using Hilbert Envelope

In the process of estimating the instantaneous pitch period from the intervals of successive positive zero crossings of the filtered signal, there could be errors due to spurious zero crossings which occur mainly if there is another impulse in between two glottal closure instants. To reduce the effects due to spurious zero crossings, the knowledge that the strength of the impulse is strongest at the GCI in each glottal cycle may be used. In order to exploit the strength of the impulses in the excitation to reduce the effects due to spurious zero crossings, the Hilbert envelope

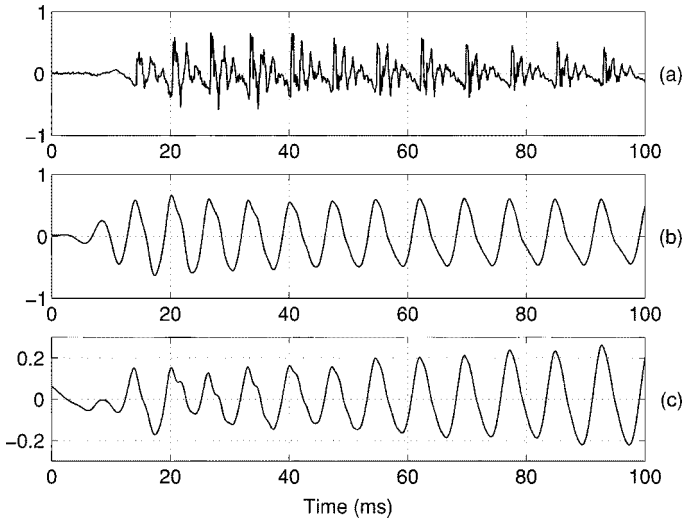

Fig. 3. (a) 100 ms segment of speech signal. Filtered signal obtained using a window length of (b) 7 ms and (c) 16 ms.

(HE) of the speech signal is computed. The HE $h[n]$ is computed from the speech signal $s[n]$ as follows:

$$h[n] = \sqrt{s^2[n] + s_h^2[n]} \tag{4}$$

where $s_h[n]$ is the Hilbert transform of $s[n]$, and is given by

$$s_h[n] = \text{IDFT}[S_h(\omega)] \tag{5}$$

where

$$S_h(\omega) = \begin{cases} +jS(\omega), & \omega < 0 \\ -jS(\omega), & \omega > 0 \end{cases} \tag{6}$$

and

$$S(\omega) = \text{DFT}[s[n]]. \tag{7}$$

Here, DFT and IDFT refer to the discrete Fourier transform and inverse discrete Fourier transform, respectively.

The Hilbert envelope contains sequence of strong impulses around the glottal closure instants, and may also contain some spurious peaks at other places due to the formant structure of the vocal-tract, and the secondary excitations in the glottal cycles, but the amplitudes of the impulses around the glottal closure instants dominate over those of the spurious impulses in the computation of the filtered signal. Hence, the filtered signal of the HE signal mainly contains the zero crossings around the instants of glottal closure. However, the zero crossings derived from the filtered signal of the HE deviate slightly (around 0.5 to 1 ms) from the actual locations of the instants of the glottal closure. In other words, the zero crossings derived from the filtered signal of the HE are not as accurate as those derived from the filtered signal of the speech signal. Hence, the accuracy of the zero crossings derived from the filtered signal of speech and the robustness of the zero crossings derived from the HE are combined to obtain an accurate and robust estimate of the instantaneous fundamental frequency.

The instantaneous pitch frequency contour obtained from the filtered signal of speech is used as the primary pitch contour, and the errors in the contour are corrected using the pitch contour derived from the HE of the speech signal. The pitch frequency contours are obtained from the zero crossings of the filtered signals for every 10 ms. The value of 10 ms is chosen for comparison with the results from other methods. Let $p_s[m]$ and $p_h[m]$ be the pitch frequency contours derived, respectively, from the speech signal and the HE of the speech signal. The following logic is used to correct the errors in $p_s[m]$

$$p[m] = \begin{cases} p_h[m], & \text{if } p_s[m] > 1.5p_h[m] \\ p_s[m], & \text{otherwise} \end{cases} \tag{8}$$

where m is the frame index for every 10 ms, and $p[m]$ is the corrected pitch contour. This correction reduces any errors in $p_s[m]$ due to spurious zero crossings.

The significance of using the pitch contour $p_h[m]$ to correct the errors in the contour $p_s[m]$ is illustrated in Fig. 4. The filtered signal shown in Fig. 4(c) for the speech segment in Fig. 4(a) contains spurious zero crossings around 0.1 to 0.2 s due to small values of the strength of excitation in this region. The filtered signal derived from the HE gives the correct zero crossings. The main idea of this logic is to correct the errors due to spurious zero crossings occurring in the filtered signal derived from the speech signal.

D. Steps in Computation of Instantaneous Fundamental Frequency From Speech Signals

1) Compute the difference speech signal $x[n]$.
2) Compute the average pitch period using the histogram of the pitch periods estimated from the autocorrelation of 30 ms speech segments.
3) Compute the output $y_2[n]$ of the cascade of two zero-frequency resonators.
4) Compute the filtered signal $y[n]$ from $y_2[n]$ using a window length corresponding to the average pitch period.
5) Compute the instantaneous fundamental (pitch) frequency from the positive zero crossings of the filtered signal. The locations of the positive zero crossings are given by the indices (n) for which $\text{diff}(\text{sgn}(y[n])) = 2$.
6) Obtain the pitch contour $p_s[m]$ for every 10 ms from the instantaneous pitch frequency by linearly interpolating the values from adjacent GCIs. This step is used mainly for comparison with the ground truth values, which are available at 10 ms intervals.
7) Compute the Hilbert envelope $h[n]$ of speech signal $s[n]$.
8) Compute the pitch contour $p_h[n]$ from the filtered signal of $h[n]$.
9) Replace the value in $p_s[m]$ with $p_h[m]$ whenever $p_s[m] > 1.5p_h[m]$.

Note: Normally, the trend removal operation in step 4) above needs to be applied only once, if the duration of the speech signal being processed is less than about 0.1 s. For longer (up to 30 s) durations, it may be necessary to apply this trend removal operation several (3 or more) times, due to rapid growth/decay of the output signal $y_2[n]$.

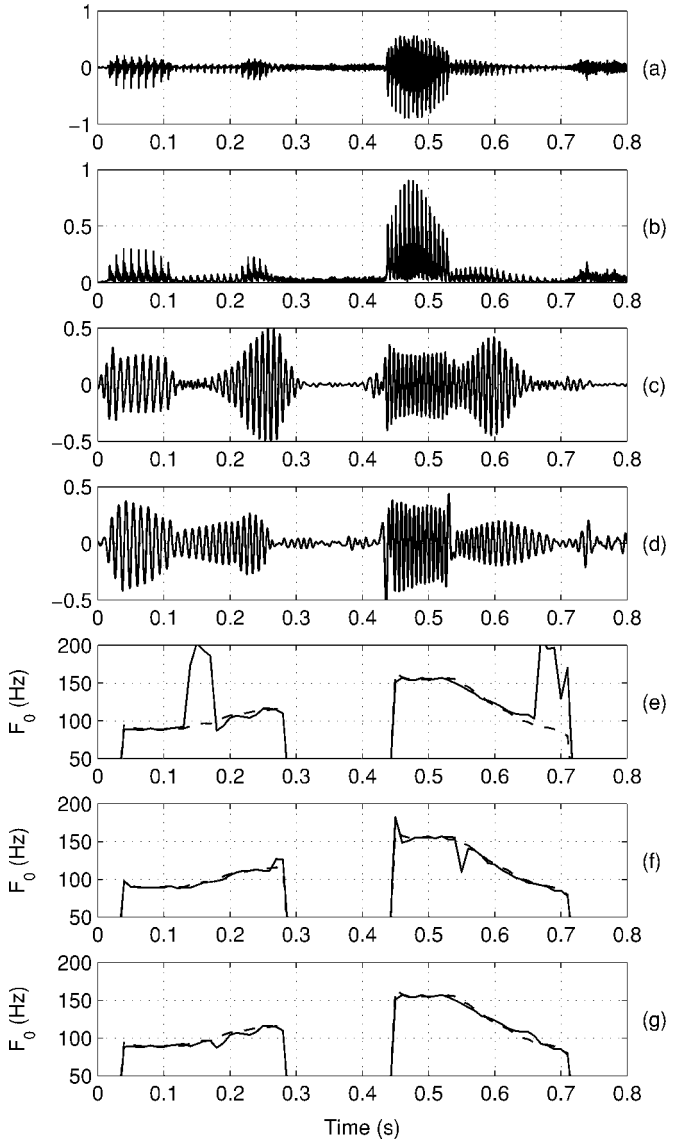

Fig. 4. (a) Speech signal. (b) Hilbert envelope. Zero frequency filtered signal derived from (c) speech signal and (d) Hilbert envelope of the speech signal. Fundamental frequency derived from (e) filtered signal of speech signal, (f) filtered signal of Hilbert envelope, and (g) correction suggested in (8). The dashed lines in the figures indicate the ground truth given by the EGG signals.

IV. Performance Evaluation and Comparison With Other Pitch Extraction Methods

In this section, the proposed method of extracting the fundamental frequency from the speech signals is compared with four existing methods in terms of accuracy in estimation and in terms of robustness against degradation. The four methods chosen for comparison are Praat's autocorrelation method [27], crosscorrelation method [28], subharmonic summation [21], and a fundamental frequency estimator (YIN) [2]. Initially the fundamental frequency estimation algorithms are evaluated on clean data. Subsequently, the robustness of the proposed method and the four existing methods are examined at different levels of degradation by white noise. A brief description of the implementation details of the four chosen methods for comparison is given below. The software codes for implementing these methods are available at the respective web sites, and are used in this study for evaluation.

A. Description of Existing Methods

Praat's Autocorrelation Method (AC) [27]: The Praat algorithm performs an acoustic periodicity detection on the basis of an accurate autocorrelation method. This method is more accurate and robust than the cepstrum-based methods and the original autocorrelation-based method [27]. It was pointed out that sampling and windowing the data cause problems in determining the peak corresponding to the fundamental period in the autocorrelation function. In this method, the autocorrelation of the original signal segment $r_x[\tau]$ is computed by dividing the autocorrelation of the windowed signal $r_a[\tau]$ with the autocorrelation of the window $r_w[\tau]$. That is

$$r_x[\tau] = \frac{r_a[\tau]}{r_w[\tau]}. \tag{9}$$

This correction does not let the autocorrelation function $r_x[\tau]$ taper to zero as the lag increases, which helps in identification of the peak corresponding to the fundamental period. To overcome the artifacts due to sampling, the algorithm employs a *sinc* interpolation around the local maxima. The interpolation provides an estimation of the fundamental period. The software code for implementation of this algorithm is available at http://www.fon.hum.uva.nl/praat/[29].

Crosscorrelation method (CC) [28]: In the computation of the autocorrelation function, fewer samples are included as the lag increases. This effect can be seen as the roll-off of the autocorrelation values for the higher lags. The values of the autocorrelation function at higher lags are important, especially for low-pitched male voices. For a 50-Hz pitch, the lag between successive pitch pulses is 200 samples at a sampling frequency of 10 kHz. To overcome this limitation in the computation of the autocorrelation function, a cross-correlation function which operates on two different data windows is used. Each value of the cross-correlation function is computed over the same number of samples. A software implementation of this algorithm is available with the Praat system [29].

Subharmonic summation (SHS) [21]: Subharmonic summation performs pitch analysis based on a spectral compression model. Since a compression on a linear scale corresponds to a shift on a logarithmic scale, the spectral compression along the linear frequency abscissa can be substituted by shifts along the logarithmic frequency abscissa. This model is equivalent to the concept that each spectral component activates not only those elements of the central pitch processor, but also those elements that have a lower harmonic relation with this component. For this reason, this method is referred to as the subharmonic summation method. The contributions of various components add up, and the activation is highest for that frequency sensitive element that is most activated by its harmonics. Hence, the maxima of the resulting sum spectrum gives an estimate of the fundamental frequency. A software implementation of this algorithm is available with the Praat system [29].

Fundamental Frequency Estimator YIN [2]: The fundamental frequency estimator YIN [2] was developed by de

Cheveigne and Kawahara, is named after the oriental yin-yang philosophical principle of balance. In this algorithm, the authors attempt to balance between autocorrelation and cancellation of the secondary peaks due to harmonics. The difficulty with autocorrelation-based methods is that the peaks occur at multiples of the fundamental period also, and it is sometimes difficult to determine which peak corresponds to the true fundamental period. YIN attempts to solve these problems in several ways. YIN is based on a difference function, which attempts to minimize the difference between the waveform and its delayed duplicate instead of maximizing the product as in the autocorrelation. The difference function is given by

$$d[\tau] = \sum_{n=1}^{N} (s[n] - s[n+\tau])^2. \tag{10}$$

In order to reduce the occurrence of subharmonic errors, YIN employs a cumulative mean function which deemphasizes higher period valleys in the difference function. The cumulative mean function is given by

$$\hat{d}[\tau] = \begin{cases} 1, & \tau = 0 \\ \frac{d[\tau]}{\frac{1}{\tau}\sum_{k=1}^{\tau} d[k]}, & \text{otherwise.} \end{cases} \tag{11}$$

The YIN method also employs a parabolic interpolation of the local minima, which has the effect of reducing the errors when the estimated pitch period is not a factor of the window length. The Matlab code for implementation of this algorithm is available at http://www.auditory.org/postings/2002/26.html [30].

B. Databases for Evaluation

Keele Database: The Keele pitch extraction reference database [31], [32] is used to evaluate the proposed method, and to compare with the existing methods. The database includes five male and five female speakers, each speaking a short story of about 35-s duration. All the speech signals were sampled at a rate of 20 kHz. This database provides a reference pitch for every 10 ms, which is obtained from a simultaneously recorded laryngograph signal, and is used as the *ground truth*. Pitch values are provided at a frame rate of 100 Hz using a 25.6 ms window. Unvoiced frames are indicated with zero pitch values, and negative values are used for uncertain frames.

CSTR Database: The CSTR database [33], [34] is formed from 50 sentences, each read by one adult male and one adult female, both with non-pathological voices. The database contains approximately five minutes of speech. The speech is recorded simultaneously with a close-talking microphone and a laryngograph in an anechoic chamber. The database is biased towards utterances containing voiced fricatives, nasals, liquids, and glides. Since some of these phones are aperiodic in comparison to vowels, standard pitch estimation methods find them difficult to analyze. In this database the reference pitch values are provided at the instants of glottal closure. Using this reference, the pitch values are derived for every 10 ms, i.e., at a frame rate of 100 Hz.

C. Evaluation Procedure

The performance of the existing as well as the proposed pitch estimation algorithms are evaluated on both Keele database and CSTR database. All the signals are downsampled to 8 kHz for this evaluation. All the methods are evaluated using a search range of 40 to 600 Hz (typical pitch frequency range of human beings). The postprocessing and voicing detection mechanisms of the existing algorithms are disabled (wherever applicable) in this evaluation.

TABLE I
PERFORMANCE OF FUNDAMENTAL FREQUENCY ESTIMATION ALGORITHMS ON CLEAN DATA. $p_s[m]$ DENOTES THE PITCH CONTOUR DERIVED FROM FILTERED SPEECH SIGNAL ALONE. $p_h[m]$ DENOTES THE PITCH CONTOUR DERIVED FROM FILTERED HE ALONE. $p[m]$ DENOTES THE PITCH CONTOUR OBTAINED BY COMBINING EVIDENCES FROM $p_s[m]$ AND $p_h[m]$ (8)

	Keele Database			CSTR Database		
Method	GE (%)	M (Hz)	SD (Hz)	GE (%)	M (Hz)	SD (Hz)
AC	5.345	2.656	3.694	5.238	4.777	6.820
CC	6.891	2.201	3.371	6.818	5.108	6.730
YIN	3.219	2.165	2.906	3.073	4.922	6.584
SHS	10.774	1.868	2.398	8.938	4.108	5.864
$p_s[m]$	2.935	3.198	4.555	3.394	5.459	6.974
$p_h[m]$	5.647	4.562	6.381	4.157	5.699	6.886
$p[m]$	2.603	3.207	4.473	1.943	5.367	6.801

The accuracy of pitch estimation methods is measured according to the following criteria [1].

- *Gross Error (GE)*: It is percentage of voiced frames with an estimated F_0 value that deviates from the reference value by more than 20%.
- *Mean Error (M):* It is the mean of the absolute value of the difference between the estimated and the reference pitch values. Gross errors are not considered in this calculation.
- *Standard Deviation (SD):* It is the standard deviation of the absolute value of the difference between estimated and reference pitch values. Gross errors are not considered in this calculation.

The reference estimates as provided in the databases are used for evaluating the pitch estimation algorithms. The reference estimates are time-shifted and aligned with the estimates of each of the methods. The best alignment is determined by taking the minimum error, over a range of time-shifts, between the estimates derived from the speech signal and the ground truth [2]. This compensation for time-shift is required due to acoustic propagation delay from glottis to microphone, and/or due to the differences in the implementations of the algorithms.

The gross estimation errors, the mean errors, and the standard deviation of errors for different fundamental frequency estimation algorithms are given in Table I. In the table, the performances of pitch contours derived from $p_s[m]$ and $p_h[m]$ are also given in addition to $p[m]$. Most of the time the percentage gross errors for the proposed method are significantly lower than the percentage gross errors for other methods. Since the number of values of pitch frequency falling within 20% of the reference values is large in the proposed method (due to inclusion of difficult and low SNR segments in the correct category, thus giving low GE), the mean error and the standard deviation error are higher compared to the other methods. The results clearly demonstrate the effectiveness of the proposed method over other methods. Note that the proposed method is based on the strength

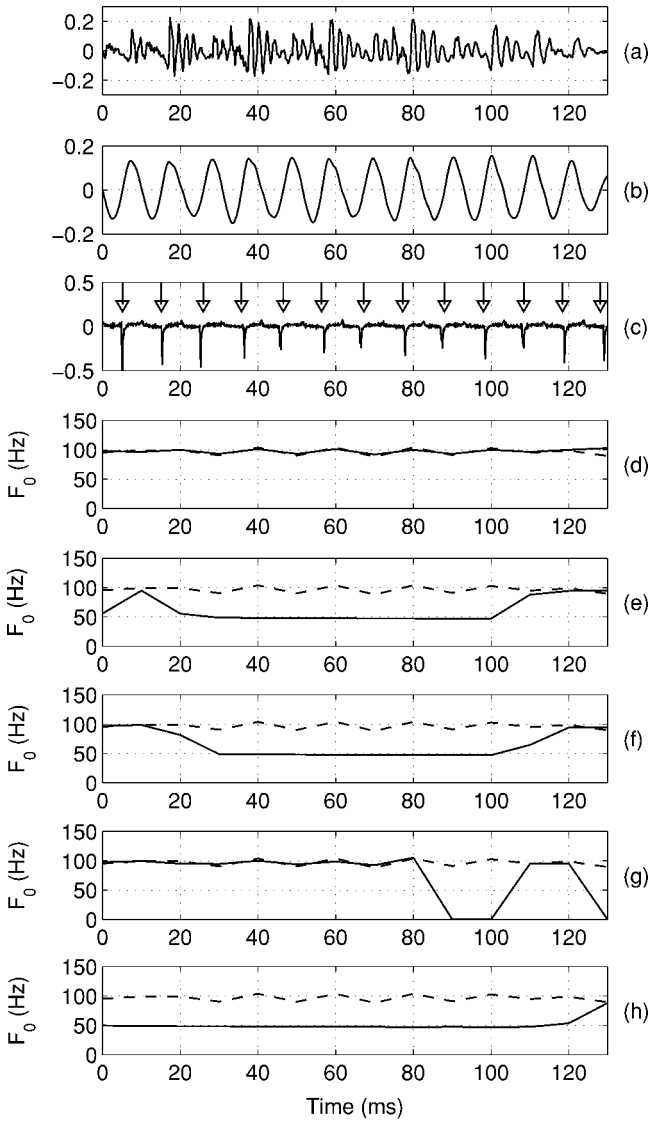

Fig. 5. (a) Speech signal. (b) Zero frequency filtered signal. (c) Differenced EGG signal. Pulses indicate the positive zero crossings of the zero-frequency filtered signal. Fundamental frequency derived from (d) proposed method, (e) Praat's autocorrelation method, (f) cross-correlation method, (g) subhormonic summation, and (h) YIN method. The dotted line corresponds to the reference pitch contour (i.e., ground truth).

of the impulse-like excitation, and it does not depend on the periodicity of the signal in successive glottal cycles. The method does not use any averaging or smoothing of the estimated values over a longer segment consisting of several glottal cycles.

The potential of the proposed method in estimating the instantaneous fundamental frequency from the speech signals is illustrated in Fig. 5. The segment of voiced speech in Fig. 5(a) is not periodic. The signal shows more similarity between alternate periods, than between adjacent periods. It is only through the analysis of the differenced EGG signal [Fig. 5(c)], the actual pitch periods could be observed. The correlation-based methods fail to estimate the actual fundamental frequency of the speech segment in these cases. On the other hand, the positive zero crossings of the filtered signal clearly show the actual glottal closure instants.

D. Evaluation Under Noisy Conditions

In this section, we study the effect of noise on the accuracy of pitch estimation algorithms. The existing methods and the proposed method were evaluated on an artificially generated noisy speech database. The noisy environment conditions were simulated by adding noise to the original speech signal at different SNRs. The noise signals were taken from Noisex-92 database [35]. Three noise environments namely white Gaussian noise, babble noise, and vehicle noise were considered in this study. The utterances were appended with silence so that the total amount of silence in each utterance is constrained to be about 60% of data, including the pauses in the utterances. The resulting data consist of about 40% speech samples, which is the amount of speech activity in a typical telephone conversation. The noise from Noisex-92 database are added to both Keele database and CSTR database to create the noisy data at SNR levels ranging from -5 to 20 dB.

Table II shows the gross estimation errors for different pitch estimation algorithms on the Keele database and CSTR database at varying levels of degradation by white noise. The performance of the correlation-based methods is similar, and is reasonable at low noise levels (up to an SNR of 10 dB). However, for higher levels of degradation, the estimation errors increase dramatically for all the systems, except for the proposed method, where the degradation in performance is somewhat gradual. Robustness of the proposed method to noise can be attributed to the impulse-like nature of the glottal closure instants in the speech signal. The energy of white noise is distributed both in time and frequency domains. While the energy of an impulse is distributed across the frequency range, and it is highly concentrated in the time domain. Therefore, the zero crossing due to an impulse is unaffected in the output of the zero-frequency resonator even in the presence of high levels of noise. Fig. 6 illustrates the robustness of the proposed method in estimating the instantaneous fundamental frequency under noisy conditions. Fig. 6(a) and (b) shows the waveforms of a weakly voiced sound under clean and degraded conditions, respectively. Fig. 6(c) and (d) shows the zero-frequency filtered signals derived from the clean [Fig. 6(a)] and the noisy signals [Fig. 6(b)], respectively. Though the individual periods can be observed from the clean signal in Fig. 6(a), it is difficult to observe any periodicity in the noisy signal shown in Fig. 6(b), but the zero crossings of the filtered signal derived from the noisy waveform remain almost the same as those derived from the clean signal, illustrating the robustness of the proposed method.

Fig. 7 illustrates the performance of the proposed method under noisy conditions, compared to the performance of the other methods. A segment of noisy speech at 0-dB SNR is shown in Fig. 7(a). The estimated pitch contour from the proposed method is given in Fig. 7(d), where the estimated values match well with the reference pitch values or ground truth (shown by dashed curves). The errors in the estimated pitch (deviation from the ground truth) can be seen clearly in all the other four methods used for comparison. Since the other methods depend mostly on the periodicity of the signal

TABLE II
GROSS ESTIMATION ERRORS (IN %) FOR DIFFERENT PITCH ESTIMATION ALGORITHMS AT VARYING LEVELS OF DEGRADATION BY WHITE NOISE

	Keele Database					CSTR Database				
SNR	AC	CC	YIN	SHS	Proposed	AC	CC	YIN	SHS	Proposed
Clean	5.345	6.891	3.219	10.774	2.603	5.238	6.818	3.073	8.938	1.943
20 dB	5.580	7.012	3.352	11.366	2.832	5.319	6.900	3.081	9.432	1.959
15 dB	5.756	7.320	3.400	12.085	3.116	5.626	7.131	3.139	9.981	2.211
10 dB	6.655	9.065	4.058	14.313	3.346	5.972	8.100	3.366	11.462	2.256
5 dB	9.173	13.462	5.955	19.562	3.907	6.249	12.287	4.933	14.868	3.069
0 dB	15.340	21.85	12.876	30.994	5.768	14.505	21.191	12.885	22.820	5.019
-5 dB	28.373	36.043	26.223	50.115	10.188	26.809	34.876	28.582	40.691	10.530

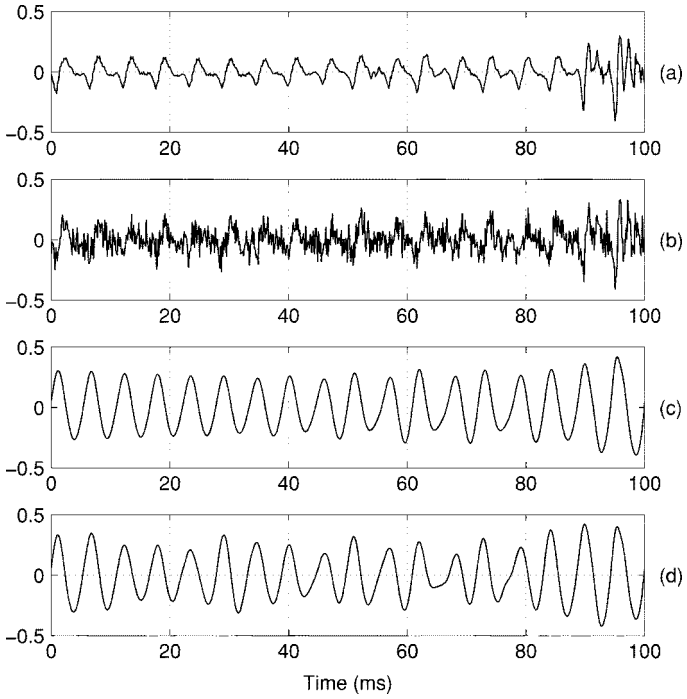

Fig. 6. (a) Speech signal of a weakly voiced sound. (b) Speech signal degraded by noise at 0-dB SNR. (c) Filtered signal derived from clean signal in (a). (d) Filtered signal derived from noisy signal in (b).

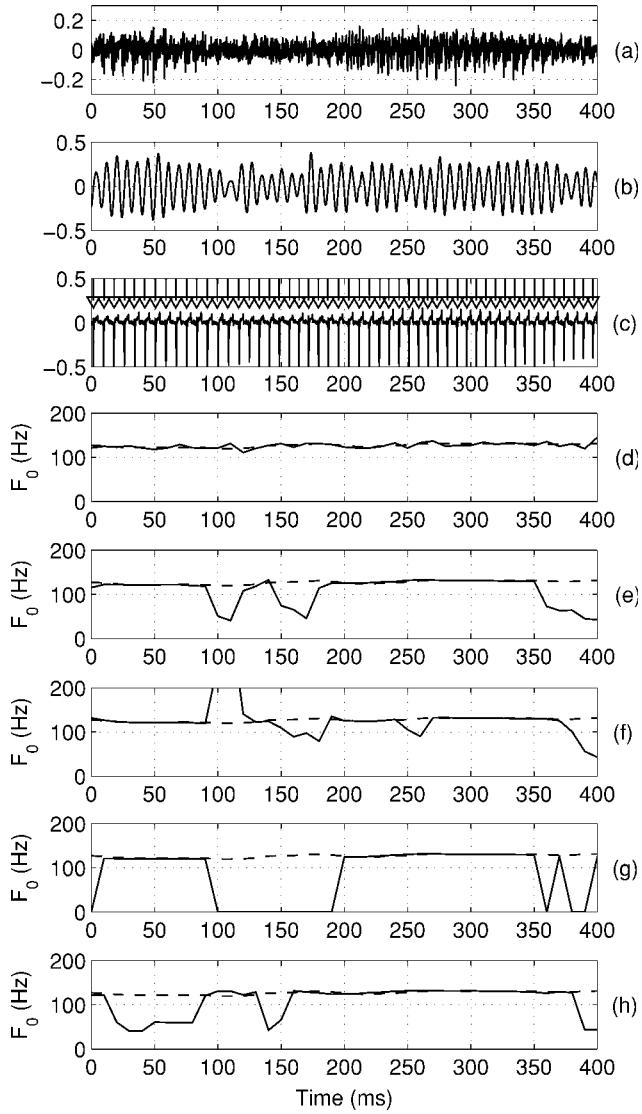

Fig. 7. (a) Speech signal at 0-dB SNR, (b) zero frequency filtered signal, (c) differenced EGG of the clean signal, pulses indicate the positive zero crossings of filtered signal in (b). F_0 derived from (d) proposed method, (e) Praat's autocorrelation method, (f) crosscorrelation method, (g) subharmonic summation, and (h) YIN method. The dotted line corresponds to the reference pitch contour.

in successive glottal cycles, the periodicity of the signal waveform is affected by noise, and hence the accuracy. Even for clean signal, there may be regions where the signal is far from periodic in successive glottal cycles, and hence there are more errors in comparison to the proposed method as can be seen in Table I. In fact, by using the additional knowledge of the strength of excitation at the impulses, it is possible to obtain the percentage gross error as low as 1.5%., but this requires significantly more heuristics which are difficult to implement automatically. Note that the proposed method does not use any knowledge of the periodicity of the speech signal, nor assume regularity of the glottal cycles. Therefore, there is scope for further improvement in the accuracy of the pitch estimation by combining the proposed method with methods based on autocorrelation.

Tables III and IV show the performance of all the five pitch estimation methods under speech-like degradation as in babble noise and low-frequency degradation as in vehicle noise. The performance of the proposed method is comparable or even

TABLE III
GROSS ESTIMATION ERRORS (IN %) FOR DIFFERENT PITCH ESTIMATION ALGORITHMS AT VARYING LEVELS OF DEGRADATION BY BABBLE NOISE

	Keele Database					CSTR Database				
SNR	AC	CC	YIN	SHS	Proposed	AC	CC	YIN	SHS	Proposed
Clean	5.345	6.891	3.219	10.774	2.603	5.238	6.818	3.073	8.938	1.943
20 dB	5.635	7.501	3.624	12.061	3.147	5.597	7.238	3.233	10.026	2.268
15 dB	6.613	8.860	4.705	13.921	3.781	6.653	8.938	3.629	11.713	2.640
10 dB	9.246	12.900	7.356	17.895	5.158	10.513	14.438	7.007	15.330	3.720
5 dB	16.155	21.579	15.745	26.35	8.618	19.438	24.400	18.947	24.177	7.205
0 dB	29.086	35.795	31.852	42.559	16.149	36.072	41.879	41.788	41.232	15.038
-5 dB	45.114	50.211	48.714	62.840	28.530	54.854	60.430	63.685	62.307	30.141

TABLE IV
GROSS ESTIMATION ERRORS (IN %) FOR DIFFERENT PITCH ESTIMATION ALGORITHMS AT VARYING LEVELS OF DEGRADATION BY VEHICLE NOISE

	Keele Database					CSTR Database				
SNR	AC	CC	YIN	SHS	Proposed	AC	CC	YIN	SHS	Proposed
Clean	5.345	6.891	3.219	10.774	2.603	5.238	6.818	3.073	8.938	1.943
20 dB	5.333	6.891	3.358	11.215	2.941	5.040	6.607	3.060	9.169	2.046
15 dB	5.550	7.428	3.708	12.067	3.104	5.069	6.372	3.184	9.701	2.281
10 dB	6.504	8.763	4.457	14.102	3.920	5.164	6.479	3.514	11.099	3.007
5 dB	9.886	13.196	7.893	18.227	6.081	6.756	8.191	5.576	14.147	5.551
0 dB	17.689	21.669	14.246	25.583	10.509	10.695	13.091	10.867	20.770	10.884
-5 dB	32.564	35.934	27.956	39.950	20.304	19.904	23.431	23.909	34.402	18.89

better than the other methods even for these two types of degradation.

V. SUMMARY AND CONCLUSION

In this paper, we have proposed a method for extracting the fundamental frequency from speech signal exploiting the impulse-like characteristic of excitation in the glottal vibrations for producing voiced speech. Since an impulse sequence has energy at all frequencies, a zero-frequency resonance filter was proposed to derive the instants of significant excitation in each glottal cycle. The method does not depend on the periodicity of glottal cycles, nor it relies on the correlation of speech signal in successive pitch periods. Thus, the method extracts the instantaneous fundamental frequency given by the reciprocal of the interval between successive glottal closure instants. Errors occur when the strength of excitation around the instant of glottal closure is not high. To correct these errors, the pitch period information derived from the zero-frequency resonator output is modified based on the pitch period information derived from the Hilbert envelope of the differenced speech signal using the proposed method. The method gives better accuracy in comparison with many standard pitch estimation algorithms. Moreover, the method was shown to be robust even under low signal-to-noise ratio conditions. Thus, the method is a very useful tool for speech analysis.

The proposed method depends only on the impulse-like excitation in each glottal cycle, and hence the intervals between successive glottal cycles are obtained without using the periodicity property in the time domain, and hence the harmonic structure in the frequency domain. Since the correlation of speech signal in successive glottal cycles is not used, the method is robust even when there are rapid changes in the successive periods of excitation, and also when there are rapid changes in the vocal-tract system, as in dynamic sounds. It may be possible to improve the performance of the proposed method by exploiting additionally the periodicity and correlation properties of the glottal cycles and speech signals, respectively.

Since the method exploits the impulse-like excitation characteristic, if there are additional impulses due to echoes or reverberation, or due to overlapping speech from a competing speaker, then the method is not likely to work well. In fact, the positive zero crossings in the filtered signal for such degraded signals may not correspond to the instants of significant excitation in the desired signal. Thus, the proposed method works well when the speech signal is captured using a close speaking microphone. For more practical degraded signals, the correlation of speech signals between adjacent glottal cycles also need to be exploited together with the proposed method.

REFERENCES

[1] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 399–418, Oct. 1976.

[2] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[3] L. Mary and B. Yegnanarayana, "Prosodic features for speaker verification," in *Proc. Interspeech'06*, Pittsburgh, PA, Sep.. 2006, pp. 917–920.

[4] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, vol. 46, no. 3–4, pp. 455–472.

[5] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Commun.*, accepted for publication.
[6] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, pp. 208–211.
[7] A. Waibel, *Prosody and Speech Recognition*. San Mateo, CA: Morgan Kaufmann, 1988.
[8] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 2037–2040.
[9] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1–2, pp. 127–154.
[10] G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Computat. Linguist.*, vol. 27, no. 1, pp. 31–57.
[11] W. J. Hess, *Pitch Determination of Speech Signals*. Berlin: Spinger-Verlag, 1983.
[12] W. J. Hess, , S. Furui and M. M. Sondhi, Eds., "Pitch and voicing determination," in *Advances in Speech Signal Processing*. New York: Marcel Dekker, 1992, pp. 3–48.
[13] D. J. Hermes, , M. Cooke, S. Beet, and M. Crawford, Eds., "Pitch analysis," in *Visual Representations of Speech Signals*. New York: Wiley, 1993, pp. 3–25.
[14] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva, *Pitch Detection With Neural-Net Classifier*, vol. 39, no. 2, pp. 298–307, Feb. 1991.
[15] A. Khurshid and S. L. Denham, "A temporal-analysis-based pitch estimation system for noisy speech with a comparative study of performance of recent systems," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1112–1124, Sep. 2004.
[16] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum aposteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, Jan. 2004.
[17] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 917–924, Mar. 1992.
[18] P. K. Ghosh, A. Ortega, and S. Narayan, "Pitch period estimation using multipulse model and wavelet transformation," in *Proc. Interspeech'07*, Antwerp, Belgium, Aug. 2007, pp. 2761–2764.
[19] B. Resch, M. Nilsson, A. Ekman, and W. B. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 15, no. 3, pp. 813–822, Mar. 2007.
[20] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1967.
[21] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.
[22] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3690–3700, Dec. 2004.
[23] J. Lee and S.-Y. Lee, "Robust fundamental frequency estimation combining contrast enhancement and feature unbiasing," *IEEE Signal Process. Lett.*, vol. 15, pp. 521–524, 2008.
[24] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AE-20, no. 5, pp. 367–377, Dec. 1972.
[25] B. Yegnanarayana, K. S. R. Murty, and S. Rajendran, "Analysis of stop consonants in Indian languages using excitation source information in speech signal," in *Proc. ISCA-ITRW Workshop on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, Jun. 4–6, 2008, p. 20.
[26] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.
[27] P. Boersma, "Accurate short-term analysis of fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*, 1993, vol. 17, pp. 97–110.
[28] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*. Boca Raton, FL: CRC, 2000.
[29] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer (Version 5.0.10). [Online]. Available: http://www.fon.hum.uva.nl/praat/
[30] A. de Cheveigne, YIN, a Fundamental Frequency Estimator for Speech and Music. [Online]. Available: http://www.auditory.org/postings/2002/26.html
[31] F. Plante, G. F. Meyer, and W. A. Aubsworth, "A pitch extraction reference database," in *Proc. Eur. Conf. Speech Commun. (Eurospeech)*, Madrid, Spain, Sep. 1995, pp. 827–840.
[32] G. F. Meyer, "Keele Pitch Database," [Online]. Available: http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html
[33] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of FQ contours for computer and intonation teaching," in *Proc. Eur. Conf. Speech Commun. (Eurospeech)*, Berlin, Germany, Sep. 1993, pp. 1003–1006.
[34] P. Bagshaw, "Evaluating pitch determination algorithms," [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/fda/
[35] *Noisex-92*, [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Sectionl/Data/noisex.htm.

B. Yegnanarayana (M'78-SM'84) received the B.Sc. degree from Andhra University, Waltair, India, in 1961, and the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science (IISc) Bangalore, India, in 1964, 1966, and 1974, respectively.

He is a Professor and Microsoft Chair at the International Institute of Information Technology (IIIT), Hyderabad, India. Prior to joining IIIT, he was a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Madras, India, from 1980 to 2006. He was the Chairman of the Department from 1985 to 1989. He was a Visiting Associate Professor of computer science at Carnegie-Mellon University, Pittsburgh, PA, from 1977 to 1980. He was a member of the faculty at the IISc from 1966 to 1978. He has supervised 32 M.S. theses and 24 Ph.D. dissertations. His research interests are in signal processing, speech, image processing, and neural networks. He has published over 300 papers in these areas in IEEE journals and other international journals, and in the proceedings of national and international conferences. He is also the author of the book *Artificial Neural Networks* (Prentice-Hall of India, 1999).

Dr. Yegnanarayana was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2003 to 2006. He is a Fellow of the Indian National Academy of Engineering, a Fellow of the Indian National Science Academy, and a Fellow of the Indian Academy of Sciences. He was the recipient of the Third IETE Prof. S. V. C. Aiya Memorial Award in 1996. He received the Prof. S. N. Mitra Memorial Award for the year 2006 from the Indian National Academy of Engineering.

K. Sri Rama Murty received the B.Tech. degree in electronics and communications engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2002. He is currently pursuing the Ph.D. degree at the Indian Institute of Technology (IIT) Madras, Chennai, India.

His research interests include signal processing, speech analysis, blind source separation, and pattern recognition.