Performance of an Event-Based Instantaneous Fundamental Frequency Estimator for Distant Speech Signals

Guruprasad Seshadri and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—This paper proposes a method for extracting the fundamental frequency of voiced speech from distant speech signals. The method is based on the impulse-like nature of excitation in voiced speech. The characteristics of impulse-like excitation are extracted by filtering the speech signal through a cascade of resonators located at zero frequency. The resulting filtered signal preserves information specific to the fundamental frequency, in the sequence of positive-to-negative zero crossings. Also, the filtered signal is free from the effects of resonances of the vocal tract. An estimate of the fundamental frequency is derived from the short-time spectrum of the filtered signal. This estimate is used to remove spurious zero crossings in the filtered signal. The proposed method depends only on the strengths of impulse-like excitations in the direct component of distant speech signals, and not on the similarity of speech signal in successive glottal cycles. Hence, the method is robust to the effects of reverberation and noise. Performance of the method is evaluated using a database of close-speaking and distant speech signals. Experiments show that the accuracy of the proposed method is significantly higher than that of existing methods based on time-domain and frequency-domain processing.**

***Index Terms*—Distant speech, fundamental frequency, impulse-like excitation, pitch, zero-frequency filtering.**

I. Introduction

SPEECH signal collected at a distance (> 3 ft) from a speaker differs in several ways from the speech signal collected close (2–3 inches) to the mouth of the speaker. The important differences between the characteristics of speech signal collected using a distant microphone (DM) and those of the speech signal collected using a close-speaking microphone (CM) are as follows. 1) The effects of radiation at far-field are different from those at the near-field. 2) The signal-to-noise (SNR) is lower in DM speech signal due to the effects of additive background noise. 3) The reverberant component in DM speech signal is significant, due to reflections, diffuse sound, and reduction in amplitude of the direct field. 4) The DM speech signal may also be affected due to interference from the speech of other speakers. Hence, the acoustic features/parameters derived from the DM speech signal are not the same as those derived from the corresponding CM speech signal. This variation in the parameters derived from the acoustic speech signal can result in degradation in the performance of speech processing systems such as automatic speech and speaker recognition systems, which are typically designed to work for CM speech. For instance, in [1], the accuracy of speaker recognition was observed to reduce significantly, when the speaker-specific models were trained using near-field speech data, and tested using far-field speech data. By contrast, human beings can perceive speech at a distance with very little difficulty, over distances ranging from 1 ft to 20 ft. While the perception of distant speech in human beings can be attributed to binaural hearing and selective attention of human listening, it also indicates that the distant speech signal does contain linguistic and speaker-specific information. Extraction of the characteristics of speech production from distant speech signals is a challenging task. This paper addresses the issue of extraction of fundamental frequency of voiced speech from distant speech signals.

Processing of distant speech signals has gained relevance in the context of extracting information from speech data collected in a meeting room scenario [2], [3]. Speech signals collected in a meeting room need to be processed to extract higher levels of information such as structure of the meeting, topics discussed in the meeting and their summaries, the various monologues and dialogues, and certain events of significance [2]. This involves processing distant speech signals for applications such as speaker turn detection, speaker recognition, language identification, and speech recognition. While acoustic features related to the short-time spectrum of speech signal are more commonly used for these applications, the fundamental frequency and its variation have also been observed to be useful for speech analysis and applications of speech processing. Variation of the fundamental frequency with time forms an important component of speech prosody, which has been used in applications such as speaker recognition [4]–[7] and language identification [7], [8]. Variation of the fundamental frequency has been exploited, along with other suprasegmental features such as durations of phones/syllables/words, short-term energy, pause duration and syllabic rate of speech, for automatic recognition of speech [3], [9], [10]. Variation of the fundamental frequency contributes to prosodic features such as pitch accent, intonational phrase boundary and prominence of speech sounds. These features have been employed to develop prosody-dependent acoustic and language models, which have helped in improving the performance of

Manuscript received November 26, 2009; revised May 29, 2010; accepted December 13, 2010. Date of publication December 23, 2010; date of current version July 15, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

G. Seshadri is with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600036, India (e-mail: gurus@cse.iitm.ac.in).

B. Yegnanarayana is with the International Institute of Information Technology Hyderabad, Hyderabad 500032, India (e-mail: yegna@iiit.ac.in).

Digital Object Identifier 10.1109/TASL.2010.2101595

conventional automatic speech recognition systems [11]–[13]. Analysis of the fundamental frequency and its variation is also useful for the detection of higher levels of information embedded in speech, such as emotional state of the speaker [14] and prominence of speech sounds and words [15].

Several algorithms have been reported in the literature for estimating the fundamental frequency in voiced speech [16], [17]. Most of these algorithms exploit the similarity of speech signal waveforms in successive glottal cycles. In the autocorrelation sequence of short segments (20–30 ms) of voiced speech, a prominent peak corresponding to the pitch period can be observed [16]–[19]. However, spurious peaks may appear due to formant structure of the vocal-tract system, or due to the position and duration of the analysis window, or due to noise. The information of the fundamental frequency and its harmonics is also reflected in the short-time spectrum of voiced speech. The fundamental frequency can be measured from the short-time spectrum, or from a nonlinearly transformed version of the short-time spectrum [20], [21]. The information of the fundamental frequency appears as rapidly varying component in the short-time spectrum, while that of the vocal-tract system appears as the gross envelope of the short-time spectrum. These two components can be separated, and pitch period can be estimated by computing cepstrum from the short-time spectrum [22]. Some algorithms remove the effect of resonances of the vocal-tract system, and depend only on the similarity of intervals of successive glottal cycles [23], [24]. Event-based approaches estimate the pitch period by measuring the time interval between two successive instants of glottal closure [25]–[28].

In the methods described above, the accuracy of the method for estimation of the fundamental frequency is evaluated using the results obtained from close-speaking (clean) speech signals and clean speech signals corrupted by additive noise. But the addition of noise to a clean speech signal can not model the characteristics of distant speech signals. Some methods for extraction of fundamental frequency from distant speech signals have been studied in [29] and [30]. In [29], pitch period is estimated using the autocorrelation of Hilbert envelope of linear prediction residual. In [30], extraction of fundamental frequency is based on the computation of cepstrum to obtain a set of candidates for the fundamental frequency, followed by a dynamic programming algorithm to select the best fundamental frequency curve. A comparison of different algorithms for estimation of fundamental frequency shows that the performance degrades severely with distance [30].

This paper proposes an algorithm for the extraction of the fundamental frequency from distant speech signals, which is based on the robustness of the impulse-like excitation in voiced speech. The key idea is that the locations of the impulse-like excitations are preserved even in distant speech signals. The proposed algorithm is based on filtering the speech signal through a zero-frequency (0 Hz) resonator. The output of the resonator preserves the fundamental frequency information, and deemphasizes the effect of resonances of the vocal tract system. The accuracy and robustness of the zero-frequency filtering of speech signals was demonstrated in the context of epoch extraction [31] and extraction of instantaneous fundamental frequency [28].

This paper is organized as follows. Section II describes the database of distant speech signals used in this study and the measures used to evaluate the performance of different methods for extraction of the fundamental frequency. The proposed method for extracting the fundamental frequency is discussed in Section III. In Section IV, the performance of the proposed method is compared with those of some existing methods. The effect of SNR of voiced segments of speech signal on the performance of the different methods is also analyzed. Conclusions are given in Section V.

II. Speech Data and Performance Measures

Speech signals from SPEECON database are used for evaluating the performance of pitch extraction algorithms [32]. The signals were collected in three different environments, namely, car interior, office, and living rooms (denoted by "public"). The signals were collected simultaneously using a close-speaking microphone, a microphone placed just below the chin of the speaker, and microphones placed at distances of 1 m and 2–3 m from the speaker. These four cases are denoted by C_0, C_1, C_2, and C_3, respectively. The noise present in the car recordings is of both stationary (engine) and instantaneous (wiper) nature [29]. Speech signals collected in the office environment are affected by stationary and white noises generated by computer fans and air-conditioning devices. Speech signals collected in living rooms are affected by babble noise and music (due to radio or television sets). Of these, reverberations are mostly present in the office and the living room environments. The reverberation time (T_{60} measure) estimated in those environments varied from 250 ms to 1.2 s. The SNR measured at the close-talking microphone is around 30 dB, while that measured at a distance of 2–3 m is around 0–5 dB.

The database consists of speech signals collected from 30 male and 30 female speakers. For each speaker, 17 utterances were recorded, resulting in about one minute of speech data per speaker. The pitch period varies from 2.5 to 15 ms, over the 1020 utterances of the 60 speakers. All speech signals were sampled at 16 kHz. The database also consists of the reference values of voiced/unvoiced decision and fundamental frequency. For each utterance, the reference values are marked once for every 10 ms, i.e., at a rate of 100 frames per second. Gross error (GE) is used to evaluate the accuracy of extraction of fundamental frequency. It is defined as the percentage of voiced frames for which the extracted value of the fundamental frequency deviates from the reference value by more than 20%. In addition, the mean (M) and standard deviation (SD) of the absolute value of the difference between the extracted and the reference values of fundamental frequency are also used for evaluation. Both M and SD are computed only for those voiced frames for which the extracted values of fundamental frequency do not deviate from the corresponding reference values by more than 20%.

III. Algorithm for Extraction of Fundamental Frequency

In [28], a method was proposed for extraction of instantaneous fundamental frequency from speech signals, by

exploiting the idea of filtering speech signal through a zero-frequency (0 Hz) resonator. The above method is applicable to clean speech signals and speech signals corrupted by additive noise. In this section, we propose some modifications to the above method, in order to extract the fundamental frequency from distant speech signals. In particular, the zero-frequency filtered signal is processed further to highlight the contribution due to the fundamental frequency.

A. Computation of Zero-Frequency Filtered Signal

The proposed method is based on filtering the speech signal through a zero-frequency (0 Hz) resonator [31]. The basis for this filtering is that the information of the discontinuities due to a sequence of impulse-like excitations is reflected across all the frequencies, including the zero frequency. The characteristics of the vocal tract system are prominent in the higher frequencies (> 300 Hz). Hence, at zero frequency (or 0 Hz), the contribution of impulse-like excitations is significant compared to that of the response of the vocal tract system. This forms the basis for filtering speech signal through a cascade of resonators, whose center frequencies are located at 0 Hz. The system function of the cascade of resonators is given by $G(z) = H(z)H(z)$, where

$$H(z) = \frac{1}{1 - 2z^{-1} + z^{-2}}. \tag{1}$$

The frequency response of $G(z)$ provides a roll-off of 24 dB per octave, and also, it does not contain any spectral nulls. The roll-off of 24 dB per octave in the magnitude response of $G(z)$ significantly reduces the influence of the resonances of the vocal tract system. The steps involved in the zero-frequency filtering are briefly summarized as follows [31].

1) Speech signals are upsampled from 16 kHz to 32 kHz. Let $s[n]$ denote the upsampled speech signal. The differenced speech signal (which helps in removing any low-frequency bias during recording) is given by $s_d[n] = s[n] - s[n-1]$.
2) The output of the cascade of resonators is given by $x[n] = s_d[n] * g[n]$, where $g[n]$ is the impulse response corresponding to the system function $G(z)$, and the symbol "$*$" denotes convolution operator. Since the poles of the system function $G(z)$ lie on the unit circle ($|z| = 1$), the output $x[n]$ has nearly polynomial growth/decay with time. An example of this trend is shown in Fig. 1(b), for the speech signal shown in Fig. 1(a).
3) The sequence of impulse-like excitations contributes small fluctuations on this growing/decaying function of time. These fluctuations can be emphasized by subtracting the local mean from the output signal $x[n]$. The length of the window chosen to compute the local mean depends on the time interval between successive impulse-like excitations. A relatively smaller window length may introduce spurious zero crossings in the filtered signal $y[n]$, while larger window length may lead to loss of some genuine zero crossings in $y[n]$. It is observed that the choice of W is not critical as long as it is in the range of 0.5 to 2 times the average pitch period of the utterance [28]. Subtraction of local mean from the signal $x[n]$ results in a signal $y[n]$, called *filtered signal* [see Fig. 1(c)], which is given by

$$y[n] = x[n] - \frac{1}{2N+1} \sum_{k=-N}^{N} x[n+k]. \tag{2}$$

Here $W = 2N+1$ is the size (in samples) of the window over which the local mean is computed. The robustness of the choice of the window length for distant speech signals is discussed in Section III-C. While the output $x[n]$ of the cascade of resonators [Fig. 1(b)] does not show the fluctuations due to impulse-like excitations, the fluctuations are clearly observed in the filtered signal $y[n]$ [Fig. 1(c)].
4) The locations of positive-to-negative zero crossings (PNZCs) of the filtered signal $y[n]$ are detected, based on the polarity of the samples of $y[n]$. These PNZCs correspond to the instants of glottal closure (i.e., epochs) in voiced speech [31]. The time interval between successive PNZCs is hypothesized as the pitch period, and the reciprocal of the time interval is hypothesized as the fundamental frequency. The upsampling of speech signal in step 1) is performed to improve the accuracy in the detection of PNZCs. The contour of the extracted fundamental frequency is shown in Fig. 1(d), along with that of the true values of the fundamental frequency.

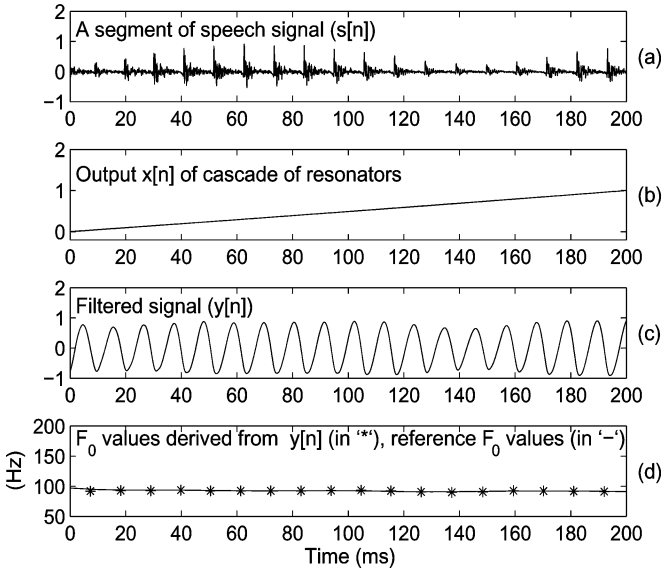


Fig. 1. Computation of zero-frequency filtered signal. (a) A segment of speech signal $s[n]$. (b) Output $x[n]$ of the cascade of resonators. (c) Filtered signal $y[n]$. (d) Contour of the fundamental frequency obtained from the filtered signal (shown using "$*$"). The solid line in (d) shows the contour of the reference values of the fundamental frequency.

B. Effect of Reverberation and Noise

The impulse-like excitations in speech signal lend robustness to the method of detecting the epochs, and hence the instantaneous fundamental frequency, even in the presence of additive noise degradations as shown in [28]. In distant speech, the degradation is not only due to additive background noise, but also due to mild reverberation. The reverberation may introduce additional impulses in the signal, but their contribution will be

less significant compared to that of the impulses due to direct component of the speech signal, because of relatively lower amplitudes of the impulses in the reflected signals.

To examine the robustness of filtering distant speech signal through a zero-frequency resonator, we consider the following cases of degradation (reverberation and noise) in signals.

1) An impulse train $u_1[n]$ which consists of a sequence of uniformly spaced impulses with a pitch period of 8 ms (fundamental frequency of 125 Hz).
2) The effect of reverberation can be modeled as the convolution of $u_1[n]$ and the room impulse response $h_r[n]$ of the ambient environment. The ambient noise is assumed to be additive and Gaussian in nature. The signal $u_2[n]$ collected at a distance can be expressed as $u_2[n] = u_1[n] * h_r[n] + v[n]$, where $v[n]$ denotes the Gaussian white noise. The energy of $v[n]$ is chosen such that the overall SNR of $u_2[n]$ is 0 dB.

We have considered a sequence of impulses as input, since it is a model of the source of excitation in the case of voiced speech. The system function $G(z)$ of the cascade of zero-frequency resonators provides a 24-dB roll-off per octave. This roll-off significantly reduces the effect of resonances of the vocal tract, which are located above 300 Hz. Hence, the filtered signal essentially represents the characteristics of the impulse-like excitations.

The room impulse response $h_r[n]$ is generated using the image method proposed by Allen and Berkley (1979) [33]. The method considers factors such as reflection order, room dimension and microphone directivity, for generating the room impulse response at a given location in the room. The following parameters were chosen for generating the room impulse response:

room dimensions: 8 m $\times$ 5 m $\times$ 3 m;
source position $[x \quad y \quad z](m) : [5 \quad 3 \quad 1.5]$;
microphone position $[x \quad y \quad z](m) : [3 \quad 1 \quad 1]$;
distance between source and microphone: 2.87 m;
reverberation time: 0.5 s;
length of impulse response: 4096 samples (at 8000 samples/s);
type of microphone: Omnidirectional;
speed of sound: 340 m/s;
sampling frequency: 8000 samples/s;
all surfaces are assumed to be fully reflective.

An implementation of the image method is obtained from the following website: http://home.tiscali.nl/ehabets/rir_generator.html

The signals $u_1[n]$ and $u_2[n]$ are filtered through the cascade of zero-frequency resonators. The resulting filtered signals, denoted by $y_1[n]$ and $y_2[n]$, respectively, are used to extract the fundamental frequency. Fig. 2 shows the clean $(u_1[n])$ and the reverberant $(u_2[n])$ signals, and the corresponding filtered signals. Fig. 2(d) shows that $y_2[n]$ consists of some spurious PNZCs compared to $y_1[n]$. The positive-to-negative zero crossings (PNZCs) derived from the filtered signals $y_1[n]$ and $y_2[n]$ are shown in Fig. 2(e). The PNZCs in Fig. 2(e) are marked by positive and negative stems for $y_1[n]$ and $y_2[n]$, respectively. Also, Fig. 2(e) shows that there is a small shift in the locations of PNZCs of $y_2[n]$, relative to those of $y_1[n]$. This shift is reflected in a small change in the fundamental frequency

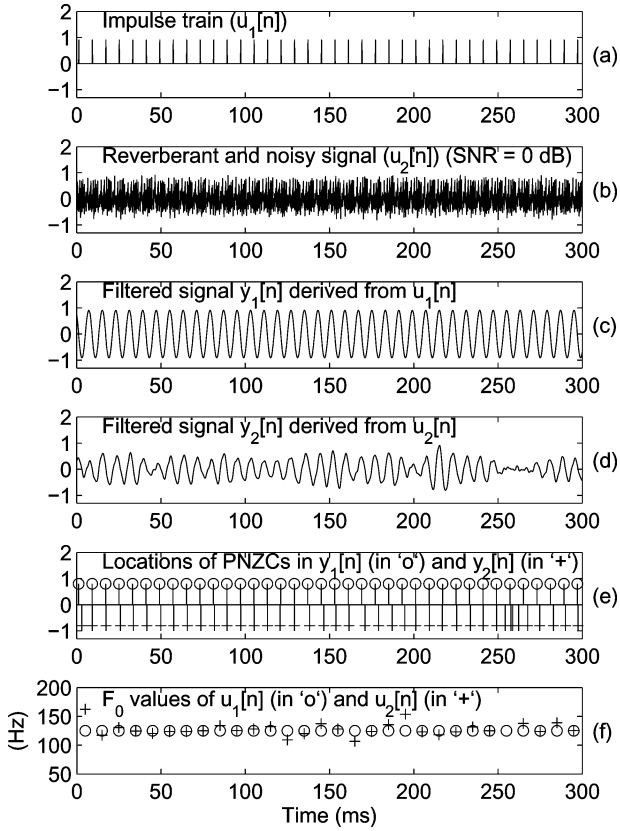


Fig. 2. Application of zero-frequency filtering on reverberant and noisy signal. (a) Sequence $u_1[n]$ of uniformly spaced impulses. (b) Signal $u_2[n]$ corrupted by reverberation and additive noise (overall SNR of 0 dB). (c) and (d) show the filtered signals $y_1[n]$ and $y_2[n]$, respectively. (e) The locations of PNZCs in $y_1[n]$ and $y_2[n]$ are indicated by positive and negative stems, respectively. (f) The values of fundamental frequency extracted from $y_1[n]$ and $y_2[n]$ are shown using the symbols "o" and "+," respectively.

of $u_2[n]$ relative to that of $u_1[n]$, which can be observed from Fig. 2(f). The gross error in the extraction of fundamental frequency in this case is 8%, computed from 400 values of the fundamental frequency over a duration of 4 s.

The above example based on the synthetic signals is given to demonstrate the robustness of zero-frequency filtering for reverberant and noisy signals. The key idea is that locations of PNZCs in the filtered signal are dictated by the strengths of the impulses in the collected signal. Even in the case of the signal collected at about 3 m from the source, the impulses due to the direct component help in preserving the locations of PNZCs in the filtered signal. In practice, the sound intensity level decreases by 6 dB per doubling of distance from the source. Let us assume a sound intensity level of 40 dB at a distance of 10 cm in front of the mouth/lips of the speaker. Then the sound intensity level of the direct component at a distance of 3 m from the speaker (along the line of sight) is about 10 dB. By contrast, the reflected components travel a greater distance (depending upon the positions of the reflecting surfaces). Thus, the direct component of speech, typically, is more dominant in the collected signal than the reflected components. This direct component also helps in preserving the phase of the original signal. It is the presence of impulse-like degradations in the noise, rather than the spectrum of the noise, which affects the performance of zero-frequency filtering method. The performance of the method degrades, if

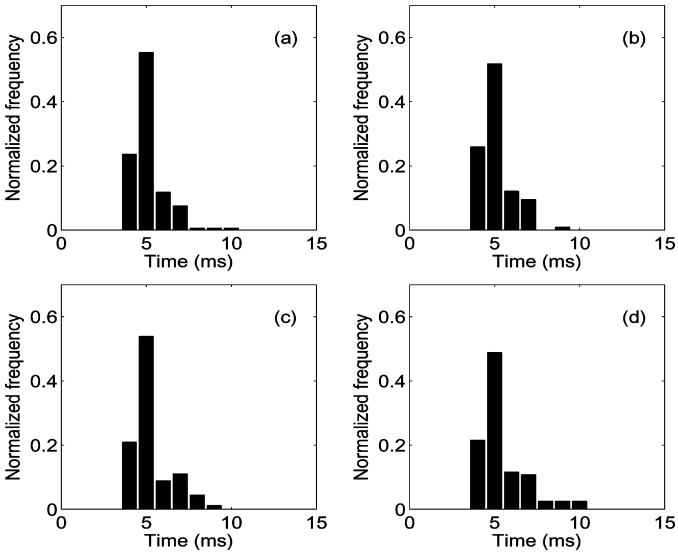


Fig. 3. Distribution of pitch periods for an utterance spoken by a female speaker, for close-speaking speech signal and for the corresponding distant speech signals. The histograms in (a)–(d) correspond to the speech signals collected over the channels C_0, C_1, C_2, and C_3, respectively.

the impulse-like degradations introduced by reverberation and noise have strengths comparable to those of the impulse-like excitations due to the direct component of the collected signal. In particular, it is the strengths of impulse-like excitations in the direct component of the signal, relative to those in the reflected components and in noise, that determine the robustness of the method.

C. Choice of Window Length

As discussed in Section III-A, the window length W for computation of local mean of $x[n]$ can be chosen in the range of 0.5 to 2 times the average pitch period of the utterance. Here we discuss the accuracy and robustness of extraction of the average value of pitch period from distant speech signals. First, the pitch period is estimated using autocorrelation sequence of the speech signal. The autocorrelation sequence is computed from 30 ms segments of speech signal for different overlapping segments, corresponding to a frame rate of 100 frames per second. For each segment, the location of the strongest peak in the autocorrelation sequence in the interval 2–15 ms is hypothesized as the pitch period. The histogram of the estimated values of the pitch period is plotted, and the pitch period corresponding to the peak in the histogram is identified. This pitch period is chosen as the average value of the pitch period of the utterance. Fig. 3 shows the histograms of the estimated values of the pitch period for an utterance spoken by a female speaker, for close-speaking speech signal and for the corresponding distant speech signals. The pitch period corresponding to the peak in the histogram is same for the close-speaking speech signal [Fig. 3(a)] and for the distant speech signals [Fig. 3(b)–(d)]. The histograms are obtained for close-speaking and distant speech signals, for all the 1020 utterances in the SPEECON database. It was observed that in the case of more than 95% of the utterances, the peaks in the histograms obtained from the distant signals lay within a deviation of 25% relative to the peaks in the histograms obtained from the corresponding close-speaking signals. Based on this observation, the pitch period $\hat{T}_0$ corresponding to the peak in the histogram is used as the average value of the pitch period, for the close-speaking and the distant speech signals. For computation of the filtered signal, the window length W is chosen as $1.5\hat{T}_0$. The local mean of the signal $x[n]$ is computed over windows of length W. The filtered signal $y[n]$ is obtained by subtracting the local mean from the signal $x[n]$, as described in Section III-A.

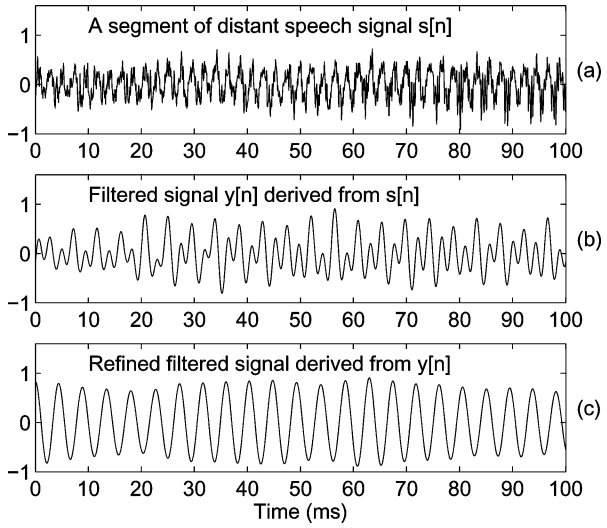


Fig. 4. (a) Speech signal $s[n]$ collected at a distance. (b) The filtered signal $y[n]$. (c) The refined filtered signal $\bar{y}[n]$.

D. Fundamental Frequency Information in the Filtered Signal

The time interval between the locations of two successive positive-to-negative zero crossings (PNZCs) in the filtered signal $y[n]$ is hypothesized as the pitch period. This time interval is an accurate estimate of the pitch period in the case of clean speech signals, and speech signals corrupted by additive noise [28]. In the case of distant speech signals, the filtered signal contains spurious PNZCs in some segments due to: 1) effect of reverberant components; 2) presence of other speakers or sound sources; and 3) noise. Fig. 4(a) shows a voiced segment of distant speech signal. The corresponding filtered signal is shown in Fig. 4(b). The information of the pitch periodicity is seen more clearly in the filtered signal [Fig. 4(b)] than in the distant speech signal [Fig. 4(a)], despite the presence of spurious zero crossings in the former.

The information specific to the fundamental frequency is also observed in the narrowband spectrogram of the filtered signal shown in Fig. 5(c). The fundamental frequency is the most dominant spectral component in the filtered signal. Also, the filtered signal is free of the effect of the formant structure. To validate this observation, the short-time spectrum of the filtered signal is computed using frames of 25 ms, and a frame shift of 5 ms. The frequency corresponding to the maximum value in the magnitude of the short-time Fourier transform (STFT) of the filtered signal is identified. This frequency is denoted by f_S, where the subscript S refers to the STFT. We first estimate the effect of the spurious zero crossings in the filtered signal $y[n]$, on the accuracy of extraction of the fundamental frequency. Let f_Z denote the reciprocal of the time interval between two successive

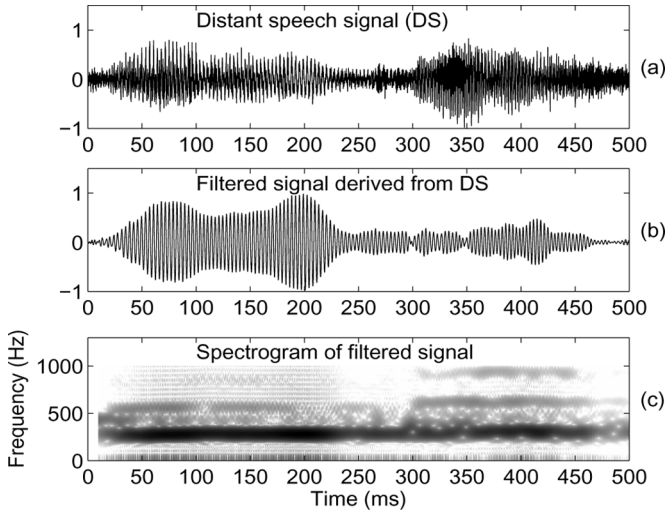


Fig. 5. Illustrating the presence of pitch information in the filtered signal. (a) Speech signal collected at a distance of 1 m. (b) The corresponding filtered signal. (c) Spectrogram of the filtered signal.

TABLE I
GROSS ERROR (IN %) IN THE EXTRACTION OF FUNDAMENTAL FREQUENCY, WHEN f_Z IS HYPOTHESIZED AS THE FUNDAMENTAL FREQUENCY

	C_0	C_1	C_2	C_3
Car	5.96	19.01	28.60	28.01
Office	5.03	8.84	13.21	24.61
Public	4.97	8.33	15.31	19.97

TABLE II
GROSS ERROR (IN %) IN THE EXTRACTION OF FUNDAMENTAL FREQUENCY, WHEN f_S IS HYPOTHESIZED AS THE FUNDAMENTAL FREQUENCY

	C_0	C_1	C_2	C_3
Car	5.85	17.62	24.67	24.41
Office	4.81	7.70	11.21	19.37
Public	4.92	7.62	13.93	15.51

PNZCs in the filtered signal $y[n]$. Here, the subscript Z indicates that the fundamental frequency is obtained using the time interval between the zero crossings of $y[n]$. Table I shows the gross error (GE) in the estimation of the fundamental frequency, when f_Z is hypothesized as the fundamental frequency. By contrast, when f_S is hypothesized as the fundamental frequency, the gross error reduces significantly as shown in Table II. This reduction in GE is observed for all the distances, and for all the three environments. Table II indicates that even for distant speech (channel C_3), f_S lies within 20% of the reference value of the fundamental frequency for about 80% of the frames in the database for office environment. This confirms the earlier observation that the fundamental frequency is the most dominant spectral component in the filtered signal, and that the effect of resonances of the vocal tract system is reduced significantly due to the zero-frequency filtering.

E. Refinement of the Filtered Signal

The spurious zero crossings in the filtered signal in some segments of voiced speech can lead to errors in the estimation of the pitch period. The knowledge of the fundamental frequency, as estimated from the filtered signal, can be exploited to remove the spurious zero crossings in the filtered signal. The short-time spectrum of the filtered signal is computed, and the frequency f_S corresponding to the largest peak value of the magnitude of STFT is obtained. The values of f_S derived from successive segments of $y[n]$ are filtered using a 5-point median filter, to eliminate erroneous values of f_S. Let $\tilde{f}_S$ denote the value obtained after the median filtering, and let $\tilde{\omega}_S$ denote the corresponding angular frequency in radians. An all-pole filter $P(z)$, whose system function is given by

$$P(z) = \frac{1}{(1 - re^{j\tilde{\omega}_S} z^{-1})(1 - re^{-j\tilde{\omega}_S} z^{-1})} \tag{3}$$

is constructed with $r = 0.99$, so as to have a sharp peak at $\tilde{\omega}_S$ in the magnitude response of $P(z)$. The signal $y[n]$ is filtered through $P(z)$, resulting in a signal $\tilde{y}[n]$, which is nearly free of the influence of spurious zero crossings in $y[n]$. Fig. 4(c) shows the output $\tilde{y}[n]$, obtained by filtering $y[n]$ through the all-pole filter $P(z)$. The spurious zero crossings present in $y[n]$ [Fig. 4(b)] have been eliminated in $\tilde{y}[n]$ [Fig. 4(c)]. Also, the pitch information is more clearly visible in $\tilde{y}[n]$, compared to $y[n]$ or the speech signal. The time interval between successive PNZCs of $\tilde{y}[n]$ is now used for deriving the fundamental frequency. Let $\tilde{f}_Z$ denote the reciprocal of the time interval between successive PNZCs in $\tilde{y}[n]$. Table III shows the gross error in the estimation of the fundamental frequency, when $\tilde{f}_Z$ is hypothesized as the fundamental frequency. The reduction in the gross error values relative to those given in Table I indicates that filtering the signal $y[n]$ through $P(z)$ does help in eliminating several spurious zero crossings in $y[n]$.

TABLE III
GROSS ERROR (IN %) IN THE EXTRACTION OF FUNDAMENTAL FREQUENCY, WHEN $\tilde{f}_Z$ IS HYPOTHESIZED AS THE FUNDAMENTAL FREQUENCY

	C_0	C_1	C_2	C_3
Car	5.77	14.15	23.68	21.70
Office	4.68	6.94	9.54	18.98
Public	4.88	6.93	11.87	14.61

The important steps in the proposed algorithm are summarized as follows.

1) Given the speech signal $s[n]$, the filtered signal $y[n]$ is computed as described in Section III-A.
2) The window length for subtraction of the local mean from the output $x[n]$ of the cascade of zero-frequency resonators is obtained as described in Section III-C.
3) The frequency f_S corresponding to the maximum value of the magnitude of short-time spectrum of the filtered signal $y[n]$ is obtained.
4) The sequence of values of f_S is filtered using a 5-point median filter, resulting in a sequence of values of $\tilde{f}_S$.
5) The all-pole filter $P(z)$ constructed using (3) is used to filter $y[n]$ to obtain $\tilde{y}[n]$.
6) The fundamental frequency is obtained as the reciprocal of the time interval between successive PNZCs of $\tilde{y}[n]$.

Note that the results in Table III show some improvement over the results in Table II. Moreover, the signal $\tilde{y}[n]$ helps to obtain

the precise locations of the zero crossings. The proposed method depends only on the sequence of impulse-like excitations, and not on the similarity of the speech signal waveform in successive glottal cycles. Hence, the method is not affected by the interaction of the fundamental frequency and the formant structure.

IV. Experimental Evaluation and Results

In this section, the performance of the proposed method for extraction of the fundamental frequency is evaluated, and compared with the results from six existing methods. These methods have been chosen as representatives of algorithms of pitch extraction in time domain and frequency domain. A brief description of the six methods is given in Section IV-A, and the performance evaluation is discussed in Section IV-B.

A. Description of Methods Used for Comparison

1) *Autocorrelation method (AC)* [34]: The short-term autocorrelation sequence of windowed speech signal is computed in such a way that it does not taper off for higher values of lag. Additionally, *sinc* interpolation is performed around the local maxima in the autocorrelation sequence to increase the accuracy of estimation. Implementation of this algorithm is available at http://www.fon.hum.uva.nl/praat/, as a part of Praat system [35].
2) *Crosscorrelation method (CC)* [36]: Crosscorrelation sequence is computed using two separate windows of the signal, which are of the same length. The objective is to overcome the effect of roll-off of the values of the autocorrelation sequence for higher lags. An implementation of this algorithm is available as a part of the Praat system [35].
3) *Robust algorithm for pitch tracking (RAPT)* [37]: In this method, peaks in the crosscorrelation sequence are identified, and the lags corresponding to these peaks are hypothesized as candidates for the pitch period. A dynamic programming algorithm is applied to select the sequence of lags that minimizes a pitch consistency cost function. An implementation of this algorithm is available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.
4) *Fundamental frequency estimator (YIN)* [38]: The autocorrelation method is modified to include the minimization of a difference function (which is the difference between a signal and its delayed version), and a parabolic interpolation of the resulting minimum. An implementation of this algorithm is available at http://audition.ens.fr/adc/sw/yin.zip.
5) *Subharmonic summation (SHS)* [20]: The algorithm is based on the observation that when a linear frequency scale is transformed to a logarithmic scale, the integral multiples of the fundamental frequency are shifted accordingly on the logarithmic scale. The abscissa f of the spectrum is transformed to a logarithmic scale $f_l = \log(f)$. For each value f_l of the frequency on the logarithmic scale, the spectral magnitudes corresponding to the various components $f_l + \log(k)$, $k = 1,2,\ldots,K$, are summed up. The value of f_l which maximizes this summation is hypothesized as the logarithm of the fundamental frequency. An implementation of this algorithm is available as part of the Praat system [35].
6) *Subharmonic-to-Harmonic ratio (SHRP)* [21]: The subharmonic-to-harmonic ratio (SHR) is defined as sum of the subharmonic amplitudes divided by the sum of the harmonic amplitudes, where the subharmonic frequencies are integral multiples of one half of the fundamental frequency. The fundamental frequency is computed on the basis of the position of global maximum of SHR, and that of a relative local maximum of SHR. An implementation of this algorithm is available at http://read.pudn.com/downloads137/sourcecode/speech/584345/shrp.m_.htm.

All the methods are evaluated using a search range of 40–600 Hz. The voicing detection mechanisms of the methods are disabled (wherever applicable) in the evaluation. The values of fundamental frequency are extracted once for every 10 ms.

B. Performance Evaluation

The performance of the proposed method and those of the six existing methods of extracting the fundamental frequency are shown in Table IV, for distant speech signals collected in three different environments. The entries against "Method-I" denote the performance of the proposed method. For each distance, the least values of gross error, mean error and standard deviation among the different methods are indicated in boldface. The gross error for the proposed method is lower than that of the other methods, for C_0, C_1, C_2, and C_3. The mean error and the standard deviation are relatively higher for the proposed method, particularly for distant speech (C_2 and C_3). This is because, the number of values of the fundamental frequency falling within 20% of the corresponding reference values is higher in the proposed method, due to inclusion of higher percentage of low SNR segments. Such segments contribute to an increase in the mean error (M) and the standard deviation (SD). Also, in the proposed method (Method-I), the fundamental frequency is obtained as the reciprocal of the time interval T_Z between two successive PNZCs of the filtered signal. Even a small perturbation in this time interval results in an increase in M and SD. For instance, in a segment with pitch period of 5 ms, an error of 0.1 ms results in an error of 4 Hz in the extraction of the fundamental frequency. To improve the accuracy of the proposed method, correlation between successive glottal cycles of speech signal can be exploited. The autocorrelation sequence is computed using 30 ms segments of speech signal. The location of the strongest peak in the autocorrelation sequence, lying within an interval of ± 1 ms from T_Z, is hypothesized as the pitch period, and its reciprocal is hypothesized as the fundamental frequency. This approach is denoted by Method-II in Table IV. The entries for Method-II indicate that the use of the autocorrelation sequence reduces the mean error significantly, although the gross error changes only marginally.

1) Variation of Gross Error With Tolerance: In Section II, gross error (GE) is defined as the percentage of voiced frames for which the extracted value of the fundamental frequency deviates from the reference value by more than 20%. The gross error can be computed by varying the percentage deviation (or

TABLE IV
PERFORMANCE OF THE PROPOSED METHOD FOR EXTRACTION OF FUNDAMENTAL FREQUENCY FROM CLOSE-SPEAKING AND DISTANT SPEECH SIGNALS. PERFORMANCES OF SIX EXISTING METHODS ARE ALSO LISTED FOR COMPARISON. FOR EACH DISTANCE, THE LEAST VALUES OF GROSS ERROR, MEAN ERROR AND STANDARD DEVIATION AMONG THE DIFFERENT METHODS ARE INDICATED IN BOLDFACE. SPEECH SIGNALS WERE COLLECTED IN THREE DIFFERENT ENVIRONMENTS

		GE (%)				M (Hz)				SD (Hz)			
	Method	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3
Car	AC	7.96	57.61	29.67	37.18	2.75	3.48	3.20	3.85	4.32	5.37	5.60	5.98
	CC	7.79	54.98	26.76	38.06	**2.35**	3.24	3.13	3.55	4.03	5.36	5.58	5.92
	SHS	16.49	51.68	60.16	57.64	2.53	2.79	2.79	3.38	**3.70**	4.25	4.41	5.12
	SHRP	10.63	64.86	52.72	56.06	2.80	**2.56**	**2.42**	**2.83**	3.79	**3.58**	**3.68**	**4.04**
	RAPT	10.46	55.76	41.46	42.27	2.74	2.91	2.55	3.96	4.29	4.57	4.42	6.37
	YIN	9.74	65.18	40.18	47.30	3.95	4.33	4.08	4.88	4.98	5.42	5.68	6.33
	Method I	5.77	14.15	23.68	21.70	4.52	5.03	6.85	7.63	3.80	6.59	8.09	8.81
	Method II	**5.75**	**12.34**	**20.34**	**19.81**	3.53	3.91	4.57	4.65	5.11	5.39	6.43	6.42
		GE (%)				M (Hz)				SD (Hz)			
		C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3
Office	AC	6.67	8.14	10.28	32.74	3.12	3.32	3.37	6.09	4.77	5.10	5.29	7.49
	CC	6.68	8.54	10.83	33.75	**2.51**	**2.78**	3.24	7.00	**4.15**	4.61	5.09	7.99
	SHS	14.06	16.56	18.91	45.63	2.94	3.00	**2.98**	**5.14**	4.20	**4.40**	**4.53**	**6.81**
	SHRP	8.98	10.96	20.15	61.17	3.32	3.38	3.37	5.43	4.54	4.70	4.74	6.86
	RAPT	9.02	10.85	16.26	44.81	2.91	3.13	3.54	6.98	4.65	4.82	5.26	7.92
	YIN	7.03	10.25	15.65	38.55	4.50	4.46	4.31	5.56	5.65	5.72	5.76	6.91
	Method I	**4.68**	**6.94**	9.54	**18.98**	4.99	5.20	5.43	9.53	4.60	5.27	5.48	10.25
	Method II	5.34	7.02	**9.34**	19.71	3.96	3.89	4.21	6.17	5.63	5.35	5.46	7.55
		GE (%)				M (Hz)				SD (Hz)			
		C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3	C_0	C_1	C_2	C_3
Public	AC	6.39	10.64	22.28	38.17	2.84	2.94	3.19	3.93	4.34	4.56	4.93	5.87
	CC	6.16	10.97	23.30	37.69	**2.39**	**2.52**	3.18	3.76	4.03	4.26	5.02	5.90
	SHS	14.58	19.13	27.29	44.66	2.68	2.64	**2.75**	**3.15**	**3.80**	**3.81**	4.18	4.73
	SHRP	8.78	14.63	33.32	65.27	2.98	2.80	2.93	3.31	4.15	3.82	**3.95**	**4.43**
	RAPT	8.35	12.55	25.60	45.66	2.71	2.81	3.29	3.47	4.26	4.35	4.74	4.92
	YIN	6.96	13.76	28.78	48.29	4.11	4.12	4.12	4.60	5.08	5.12	5.28	5.75
	Method I	**4.88**	6.93	11.87	14.61	4.86	4.83	5.70	7.29	4.32	4.44	6.63	9.02
	Method II	5.03	**6.76**	**10.94**	**14.08**	3.65	3.50	3.98	4.59	5.02	4.47	5.09	5.83

tolerance) between the extracted and the reference values of the fundamental frequency. Fig. 6(a) shows the variation of GE for different values of the deviation, for distant speech signals (C_3). The proposed method (Method-II, denoted by M-II) has smaller values of GE compared to the other methods, even for smaller values of the deviation (3% to 10%). Also, the reduction in GE over the range 10%–30% is greater for the proposed method (M-II), compared to the other methods. Fig. 6(b) shows the variation of GE for the proposed method (M-II), for close-speaking and distant speech signals. The reduction in GE with the deviation is more pronounced for close-speaking speech (C_0), compared to the distant speech (C_3).

2) Effect of SNR and Duration of Voiced Segments: SNR of distant speech signal reduces as a function of distance, since the amplitude of the direct component reduces due to reverberation and noise. It is not always possible to estimate the SNR of the speech signals collected at a distance, due to the unknown nature of the ambiance and the nonstationary nature of the noise. Here, the short-term power of the close-speaking speech signal is used to measure the SNR, for close-speaking as well as the distant speech signals. The short-term signal power (in dB) is given by $E_c = 10\log_{10}\left\{(1/N)\sum_{n=0}^{N-1} s^2[n]\right\}$. The length N is chosen corresponding to a duration of 25 ms. The short-term power E_c is computed from close-speaking speech signal at a frame-rate of 100 frames/s. The values of the fundamental frequency are obtained for the corresponding frames, from the close-speaking and the distant speech signals. The distribution of E_c, estimated from the voiced regions of the 1020 close-speaking utterances in SPEECON database, is shown in Fig. 7(a). Fig. 7(b) shows the gross error obtained using all those voiced frames of distant speech signals (C_2), for which E_c is greater than a threshold. For instance, the gross error in Fig. 7(b) corresponding to the

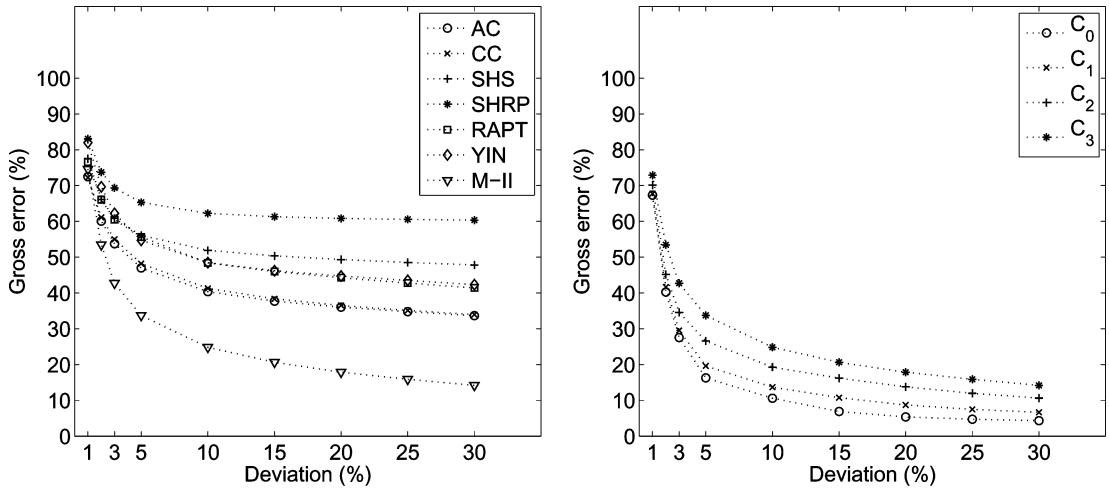


Fig. 6. Variation of gross error (GE) as a function of the deviation between extracted values and the corresponding reference values of the fundamental frequency. (a) Variation of GE for distant speech signals (collected over C_3) for the different methods. (b) Variation of GE for the proposed method (M-II), for close-speaking and distant speech signals.

abscissa of -55 dB is obtained using all those frames of distant speech signals (C_2), for which the short-term power E_c of the close-speaking signal is greater than or equal to -55 dB. Fig. 7(c) shows the variation of gross error (GE) for distant speech signals collected over C_3. Fig. 7(b) and (c) indicates that the value of GE reduces systematically when voiced segments with lower energy are progressively excluded from the computation of GE. This trend is expected, since, smaller the value of E_c of a voiced segment of the close-speaking speech signal, the greater is the degradation suffered by the corresponding segment of the distant speech signal. In Fig. 7(b) and (c), the proposed method (M-II) shows significantly lower values of GE compared to the other methods. This is attributed to the impulse-like nature of excitation, which imparts higher SNR to a short segment of speech signal in its vicinity.

While the overall SNR may be small, the SNR in short segments of speech signal around the impulse-like excitations is relatively higher, and can help in processing of speech signals in the presence of degradations. Algorithms for estimation of the fundamental frequency are also sensitive to the duration of voiced segments. It is observed that all the methods suffer greater error for voiced segments of shorter duration (50–100 ms), although this error is significantly less for the proposed method (M-II). It is likely that the voiced segments of shorter duration are also the segments with low SNR. In the remaining range of duration (100–600 ms), the gross error does not vary appreciably for different methods.

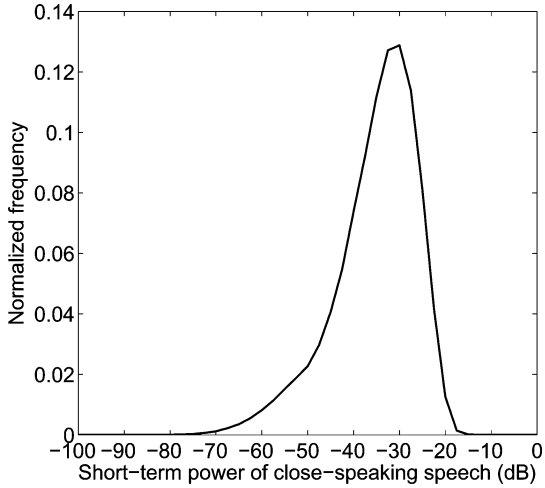
3) Significance of Regions of High SNR: Algorithms for estimation of the fundamental frequency, which depend on the similarity of speech signal in successive glottal cycles, show significant degradation in performance in the case of distant speech signals. This can be explained as follows. For a segment of speech signal $s[n]$ of L samples, the covariance sequence $c[l]$ is computed as

$$c[l] = \frac{\Sigma_{n=0}^{L-1} s[n]s[n+l]}{\sqrt{\Sigma_{n=0}^{L-1} s^2[n]}\sqrt{\Sigma_{n=0}^{L-1} s^2[n+l]}} \quad l = 0, 1, 2, \ldots, L-1 \tag{4}$$

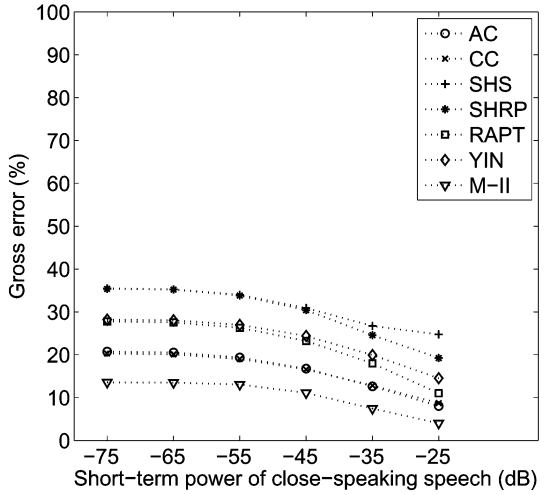
where l is the lag, and L is chosen corresponding to 20 ms. The maximum value of the covariance sequence is identified for lags that lie between 2 and 20 ms. We denote this maximum value as correlation coefficient, which is a measure of similarity of the speech signals in successive glottal cycles. The distribution of the correlation coefficient for close-speaking and distant speech signals is shown in Fig. 8. Clearly, the signal similarity in successive glottal cycles reduces as a function of distance. This reduction in signal similarity causes degradation in the performance of the algorithms of fundamental frequency estimation, which attempt to exploit the signal similarity in time or frequency domains. By contrast, the robustness of the proposed method is attributed to the characteristics of impulse-like excitations in voiced speech. Due to the impulse-like nature of excitation, short segments of speech signal around the excitations have higher SNR. Such segments are robust to the effects of noise, and help in the perception of speech even at a distance.

4) Comparison With Other Methods: A comparison of the performance of the proposed method (M-I) with that of the method proposed in [28] is given in Table V, for close-speaking speech signals (C_0). The method given in [28] performs better than the proposed method in the case of close-speaking speech signals, due to the validation of the instantaneous fundamental frequency using the Hilbert envelope of speech signal. The zero crossings of the filtered signal derived from the Hilbert envelope of speech signal are used to correct any errors in the contour of the instantaneous fundamental frequency. The method proposed in [28] is thus more suitable for extraction of fundamental frequency from close-speaking speech signals, although further processing is required for the case of distant speech signals.

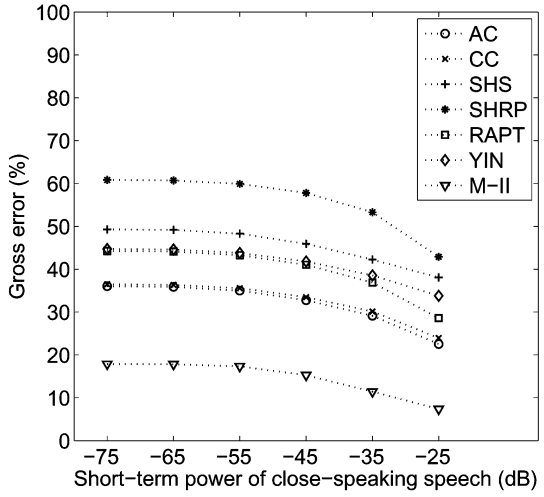
A method based on the autocorrelation of the Hilbert envelope of the LP residual of speech signal was evaluated for SPEECON database [29]. The values of gross error reported in Table I of [29] (4.78%, 8.87%, 15.70%, and 9.58% for C_0, C_1, C_2 and C_3, respectively) are significantly lower than those obtained by the proposed method. However, those values of gross



(a)



(b)



(c)

Fig. 7. (a) Distribution of short-term power E_c of voiced segments of close-speaking speech signals. Variation of gross error with E_c for distant speech signals collected over (b) C_2 and (c) C_3.

error were computed using only the detected voiced frames, which were considerably lower for distant speech [29].

The errors quoted for the detection of voiced regions for the distant speech signals were 41.91% and 66.51% for C_2 and C_3, respectively, indicating that only 58.09% and 33.49% of the voiced frames for C_2 and C_3, respectively, were used for the computation of the gross errors in [29]. Thus, voiced regions of low SNR were not included in the computation of gross error. On the other hand, the gross error in the proposed method is computed using all the voiced regions. Therefore, to compare with the results in [29], we consider the distribution of short-term power E_c of the voiced segments of close-speaking speech signals (C_0) [Fig. 7(a)]. We then identify the values of E_c, above which, 58.09% and 33.49% of the voiced regions of speech signals are retained. For these values of E_c (-34.2 dB and -29.5 dB), the gross errors from Fig. 7(b) and (c) are (approximately) 8% and 9% for C_2 and C_3, respectively, which are lower than 15.7% and 9.58% given in [29]. Note that the average gross error for the case C_1 from Table IV is 8.68% which includes all the voiced frames, whereas the average gross error for the case C_1 using 82.20% of the detected voiced frames is 8.87% in [29]. For the case C_0, the minimum average gross error is 4.2% from Table V, whereas the corresponding value from [29] is 4.78%. Thus, the values of the gross error obtained by the proposed method are either lower or comparable with the gross error reported in [29] for these four cases.

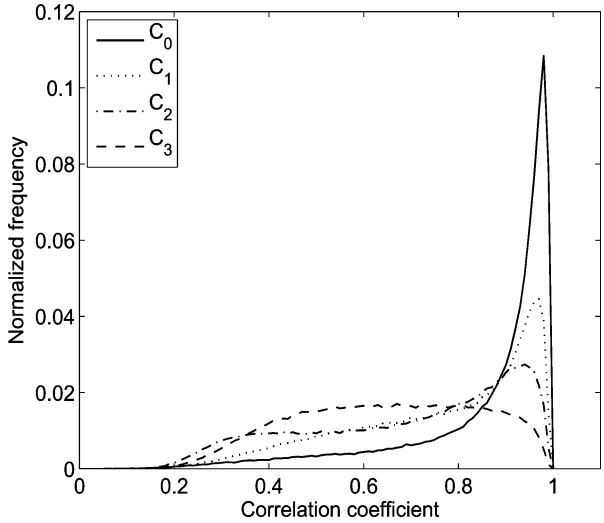


Fig. 8. Distribution of correlation coefficient for close-speaking and distant speech signals.

TABLE V
COMPARISON OF PERFORMANCE OF THE PROPOSED METHOD WITH THAT OF THE METHOD DESCRIBED IN [28], FOR CLOSE-SPEAKING SPEECH SIGNALS (C_0). GROSS ERRORS ARE INDICATED IN PERCENTAGE

	M-I	Method in [28]
Car	5.77	4.97
Office	4.68	4.21
Public	4.88	3.82

V. CONCLUSION

This paper addresses the issue of extraction of the fundamental frequency from distant speech signals. The proposed method exploits the robustness of the impulse-like excitations in voiced speech. The key idea is that the short segments of

high SNR in the speech signal located in the vicinity of the impulse-like excitations are robust to the effects of reverberation and noise. The strengths of the impulse-like excitations in the direct component are relatively stronger than those of reverberant components and noise. The locations of the impulse-like excitations in distant speech signal are derived by filtering the speech signal through a cascade of resonators located at zero frequency. The filtered signal preserves the information specific to the fundamental frequency in the form of zero crossings, and is free from the effects of resonances of the vocal tract. An estimate of the fundamental frequency derived from the filtered signal is used to remove spurious zero crossings in the filtered signal. For distant speech signals, the proposed method gives significantly better accuracy in comparison with many existing methods of estimation of the fundamental frequency. The performance of the proposed method degrades for voiced segments with very low SNR and for segments with short duration (50–100 ms), although the degradation is significantly less compared to the other methods.

There are other types of degradations for which the method proposed in this paper may not be directly applicable. For example, in the case of mixed speech from two or more speakers, the signal will have impulses of significant strengths of one speaker, interfering with those of the other speakers. Hence, the method proposed in this paper cannot be applied directly. However, it is possible to estimate the instantaneous fundamental frequency contours of individual speakers by collecting the mixed signals from two or more spatially separated microphones. In such a case, strengths of the impulses due to one speaker can be enhanced by compensating for the fixed delay of the impulse sequences of the speaker at the two microphones. The coherent addition of impulses due to one speaker enhances the impulse sequence corresponding to that speaker. At the same time, the coherent addition reduces the importance of impulse sequences corresponding to other speakers, due to incoherence of the sequences for that delay [39].

ACKNOWLEDGMENT

The authors would like to thank the members of ECESS Consortium (http://www.ecess.eu), and in particular, Dr. H. Höge of Siemens AG, Corporate Technology, Germany, for granting the use of SPEECON database along with reference pitch information.

REFERENCES

[1] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2023–2032, Sep. 2007.

[2] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic Bayesian networks," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 25–36, Jan. 2007.

[3] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 1303–1314, Sep. 2008.

[4] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stocke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, vol. 46, no. 3–4, pp. 455–472, Jul. 2005.

[5] B. Yegnanarayana, S. R. M. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 575–582, Jul. 2005.

[6] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2095–2103, Sep. 2007.

[7] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech Commun.*, vol. 50, no. 10, pp. 782–796, Oct. 2008.

[8] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *J. Acoust. Soc. Amer.*, vol. 105, no. 1, pp. 512–521, Jan. 1999.

[9] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1–2, pp. 127–154, Sep. 2000.

[10] D. H. Milone and A. J. Rubio, "Prosodic and accentual information for automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 321–333, Jul. 2003.

[11] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, "Prosody dependent speech recognition on radio news corpus of American English," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 232–245, Jan. 2006.

[12] V. K. R. Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 797–811, May 2008.

[13] S. Ananthakrishnan and S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 138–149, Jan. 2009.

[14] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 582–596, May 2009.

[15] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 5, pp. 1009–1024, Jul. 2009.

[16] W. J. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.

[17] W. J. Hess, *Pitch and Voicing Determination*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992.

[18] D. Krubsack and R. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 319–329, Feb. 1991.

[19] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 727–730, Oct. 2001.

[20] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[21] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, May 2002, pp. 333–336.

[22] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[23] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AE-20, no. 5, pp. 367–377, Dec. 1972.

[24] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Montreal, QC, Canada, Mar. 2004, vol. 1, pp. 109–112.

[25] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1805–1815, Dec. 1989.

[26] 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.

[27] 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.

[28] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614–624, May 2009.

[29] B. Kotnik, H. Höge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions," in *Proc. 3rd Int. Conf. Speech Prosody*, Dresden, Germany, May 2006, pp. 149–152.

[30] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Honululu, HI, May 2007, vol. 4, pp. 1057–1060.

[31] 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.

[32] D. Iskra, B. Grosskopf, K. Marasek, H. V. D. Heuvel, F. Diehl, and A. Kiessling, "SPEECON—Speech databases for consumer devices: Database specification and validation," in *Proc. 3rd Int. Conf. Lang. Resources Eval. (LREC)*, Las Palmas, Spain, May 2002, pp. 329–333.
[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[34] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*. Amsterdamn, The Netherlands: Univ. of Amsterdam, 1993, vol. 17, pp. 97–110.
[35] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer (Version 4.3.14) 2005 [Online]. Available: http://www.praat.org/
[36] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*. Boca Raton, FL: CRC Press, 2000.
[37] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT)*. Amsterdam, The Netherlands: Speech coding and synthesis, Elsevier Science, 1995.
[38] A. de Cheveigné 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.
[39] B. Yegnanarayana and S. R. M. Prasanna, "Analysis of instantaneous F_0 contours from two speakers mixed signal using zero frequency filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, Mar. 2010, pp. 5074–5077.

Guruprasad Seshadri received the B.E. degree in electronics and communications engineering from The National Institute of Engineering, University of Mysore, Karnataka, India, in 1998, and M.S. degree from Indian Institute of Technology (IIT) Madras, Chennai, India, in 2004. He is currently pursuing the Ph.D. degree at IIT Madras.

His research interests include speech signal processing and pattern recognition.

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. 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 Indian Institute of Science (IISc), Bangalore, from 1966 to 1978. He has supervised 36 M.S. theses and 26 Ph.D. dissertations. His research interests are in signal processing, speech, image processing, and neural networks. He has published over 350 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.