Epoch Extraction From Speech Signals

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

***Abstract*—Epoch is the instant of significant excitation of the vocal-tract system during production of speech. For most voiced speech, the most significant excitation takes place around the instant of glottal closure. Extraction of epochs from speech is a challenging task due to time-varying characteristics of the source and the system. Most epoch extraction methods attempt to remove the characteristics of the vocal-tract system, in order to emphasize the excitation characteristics in the residual. The performance of such methods depends critically on our ability to model the system. In this paper, we propose a method for epoch extraction which does not depend critically on characteristics of the time-varying vocal-tract system. The method exploits the nature of impulse-like excitation. The proposed zero resonance frequency filter output brings out the epoch locations with high accuracy and reliability. The performance of the method is demonstrated using CMU-Arctic database using the epoch information from the electro-glottograph as reference. The proposed method performs significantly better than the other methods currently available for epoch extraction. The interesting part of the results is that the epoch extraction by the proposed method seems to be robust against degradations like white noise, babble, high-frequency channel, and vehicle noise.**

***Index Terms*—Epoch extraction, glottal closure instant, group-delay, Hilbert envelope, instantaneous frequency.**

I. Introduction

THE INSTANT of significant excitation of the vocal-tract system is referred to as the epoch. An excitation is termed as significant if it is impulse-like with strength substantially larger than the strengths of impulses in the neighborhood. In the context of speech, most of the significant excitation takes place due to glottal vibration. The exceptions are strong burst releases of very short durations. During the glottal vibration, the major impulse-like excitation takes place during the closing phase of the glottal cycle, due to abrupt closure of the vocal folds. Determining the epochs from a voiced speech signal is the main objective of this paper.

A. Significance of Epochs in Speech Analysis

Voiced speech analysis consists of determining the frequency response of the vocal-tract system and the glottal pulses representing the excitation source. Although the source of excitation for voiced speech is a sequence of glottal pulses, the significant excitation of the vocal-tract system is within a glottal pulse. The significant excitation can be considered to occur at the instant of glottal closure, called the epoch. Many speech analysis situations depend on the accurate estimation of the epoch locations within a glottal pulse. For example, knowledge of the epoch locations is useful for accurate estimation of the fundamental frequency (f_0). Often the glottal airflow is zero soon after the glottal closure. As a result the supralaryngeal vocal-tract is acoustically decoupled from the trachea. Hence, the speech signal in the closed phase region represents the free resonances of the supralaryngeal vocal-tract system. Analysis of the speech signal in the closed phase regions provides an accurate estimate of the frequency response of the supralaryngeal vocal-tract system [1], [2]. With the knowledge of the epochs, it is possible to determine the characteristics of the voice source by a careful analysis of the signal within a glottal pulse. The epochs can be used as pitch markers for prosody manipulation, which is useful in applications like text-to-speech synthesis, voice conversion and speech rate conversion [3], [4]. Knowledge of the epoch locations may be used for estimating the time-delay between speech signals collected over a pair of spatially distributed microphones [5]. The segmental signal-to-noise ratio (SNR) of the speech signal is high in the regions around epochs, and hence, it is possible to enhance the speech by exploiting the characteristics of speech signals around the epochs [6]. It has been shown that the excitation features derived from the regions around the epoch locations provide complementary speaker-specific information to the existing spectral features [7], [8].

As a result of significant excitation at the epochs, the regions in the speech signal that immediately follow them are relatively more robust to (external) degradations than other regions. The instants of significant excitation play an important role in human perception also. It is because of the epochs in speech that human beings seem to be able to perceive speech even at a distance (e.g., 10 ft or more) from the source, even though the spectral components of the direct signal suffer an attenuation of around 60 dB. For example, we may not be able to get the message in whispered speech by listening to it at a distance of 10 ft or more due to absence of epochs. The neural mechanism of human beings seem to have the ability of processing selectively the robust regions around the epochs for extracting the acoustic cues even under degraded conditions. It is the ability of human beings to focus on these microlevel events that may be responsible for extracting robust and reliable speech information even under severe degradation such as noise, reverberation, presence of other speakers and channel variations.

B. Review of the Existing Methods

Normally, epochs are attributed to the glottal closure instants (GCIs) of the glottal cycles. Most epoch extraction methods rely on the error signal derived from the speech waveform after removing the predictable portion (second-order correlations). The error signal is usually derived by performing linear prediction (LP) analysis of the speech signal [9]. The energy of the error signal is computed in blocks of small interval (1–2 ms), and the point where the computed energy is maximum is hypothesized

Manuscript received April 08, 2008; revised July 04, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

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

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

Digital Object Identifier 10.1109/TASL.2008.2004526

as the instant of significant excitation. Some methods also exploit the periodicity property of the signal in the adjacent cycles for epoch extraction. The method proposed in this paper assumes and exploits the impulse-like characteristic of the excitation. The intervals between the adjacent impulses are not necessarily equal, i.e., the glottal cycles need not be periodic even in short intervals of a few (2–4) glottal cycles.

The first contribution to the detection of epochs was due to Sobakin [10]. A slightly modified version was proposed by Strube [11]. In Strube's work, some predictor methods based on LP analysis for the determination of the epochs were reviewed. These methods do not always yield reliable results. Sobakin's method using the determinant of the autocovariance matrix was examined critically, and it was shown that the determinant was maximum if the beginning of the interval, on which the autocovariance matrix was computed, coincided with the glottal closure.

In [12], a method based on the composite signal decomposition was proposed for epoch extraction of voiced speech. A superposition of nearly identical waveforms was referred to as a composite signal. The epoch filter proposed in this work, computes the Hilbert envelope of the highpass filtered composite signal to locate the epoch instants. It was shown that the instants of excitation of the vocal-tract could be identified precisely even for continuous speech. However, this method is suitable for analyzing only clean speech.

The error signal obtained in the LP analysis, referred to as the LP residual, is known to contain information pertaining to epochs. A large value of the LP residual within a pitch period is supposed to indicate the epoch location [13]. However, epoch identification directly from the LP residual is not recommended [11], because the LP residual contains peaks of random polarity around the epochs. This makes unambiguous identification of the epochs from the LP residual difficult. A detailed study was made on the determination of the epochs from the LP residual [14], and a method for unambiguous identification of epochs from the LP residual was proposed. A least-squares approach for glottal inverse filtering from the acoustic speech waveform was proposed in [15]. In this paper, covariance analysis was discussed for accurately performing the glottal inverse filtering from the acoustic speech waveform.

A method based on maximum-likelihood theory for epoch determination was proposed in [16]. In this method, the speech signal was processed to get the maximum-likelihood epoch detection (MLED) signal. The strongest positive pulse in the MLED signal indicates the epoch location within a pitch period. However, the MLED signal creates not only a strong and sharp epoch pulse, but also a set of weaker pulses which represent the suboptimal epoch candidates within a pitch period. Hence, a selection function was derived using the speech signal and its Hilbert transform, which emphasized the contrast between the epoch and the suboptimal pulses. Using the MLED signal and the selection signal with appropriate threshold, the epochs were detected. The limitation of this method is the choice of window for deriving the selection function, and also the use of threshold for deciding the epochs.

A Frobenius norm approach for detecting the epochs was proposed in [17]. In this paper, a new approach based on singular value decomposition (SVD) was proposed. The SVD method amounts to calculating the Frobenius norms of signal matrices, and is therefore, computationally efficient. The method was shown to be working only for vowel segments. No attempt was made in detecting epochs in difficult cases like nasals, voiced consonants, and semivowels.

A method for detecting the epochs in a speech signal using the properties of minimum phase signals and group-delay function was proposed in [18]. The method is based on the fact that the average value of the group-delay function of a signal within an analysis frame corresponds to the location of the significant excitation. An improved method based on the computation of the group-delay function directly from the speech signal was proposed in [19]. Robustness of the group-delay based method against additive noise and channel distortions was studied in [20]. Four measures of group-delay (average group-delay, zero frequency group-delay, energy-weighted group-delay, and energy-weighted phase) and their use for epoch detection was investigated in [21]. In this paper, the effect of the length of analysis window, the tradeoff between the detection rate and the timing error, and the computational cost of evaluating the measures were examined in detail. In this paper, it was shown that the energy-weighted measures performed better than the other two measures. A dynamic programming projected phase-slope algorithm (DYPSA) for automatic estimation of glottal closure instants in voiced speech was presented in [22] and [23]. In this method, the candidates for GCI were obtained from the zero crossings of the phase-slope function derived from the energy-weighted group-delay, and were refined by employing a dynamic programming algorithm. In this paper, it was shown that DYPSA performed better than the existing methods.

Epoch is an instant property, but, in most of the methods discussed above (except the group-delay based methods), the epochs are detected by employing block processing approaches, which result in ambiguity about the precise location of the epochs. Most of the existing methods rely on the LP residual signal derived by inverse filtering the speech signal. Though these methods work well in most cases, they need to deal with the following issues: 1) selection of parameters (order of LP analysis, length of the window) for deriving the error signal; 2) dependence of these methods on the energy of the error signal, which in turn depends on the energy of the signal; 3) the accuracy with which the epochs can be resolved decreases as a result of block processing; 4) setting a threshold value to take an unambiguous decision on the presence of an epoch; 5) though some of these methods exploit periodicity for accurate estimation of epoch locations, the excitation impulses need not be periodic. In general, it is difficult to detect the epochs in the case of low voiced consonants, nasals and semivowels, breathy voices, and female speakers.

In this paper, we propose a new method for epoch extraction which is based on the assumption that the major source of excitation of the vocal-tract system is due to a sequence of impulse-like events in the glottal vibration. The impulse excitation to the system results in a discontinuity in frequency in the output signal. We propose a novel approach to detect the location of the discontinuity in frequency in the output signal by confining the analysis around a single frequency. In Section II, we discuss the basic principle of the proposed method and illustrate the principle for several cases of synthetic excitation signals. In Section III, we discuss the issues involved in applying the method directly on speech data. In Section IV, we

present our proposed method to extract epochs from the speech signal. In Section V, the performance of the proposed method in terms of identification accuracy is given, and the results are compared with three existing methods for epoch extraction. In Section VI, the performance of the proposed method is evaluated for different types of degradations, and the results are compared with the existing methods. Finally, in Section VII we summarize the contributions of this paper, and discuss some limitations of the proposed method which prompt further investigation for extracting epochs from speech signals recorded in practical environments.

II. Basis for the Proposed Method of Epoch Extraction

Speech is produced by exciting the time-varying vocal-tract system by one or more of the following three types of excitation: 1) glottal vibration; 2) frication; 3) burst. The primary mode of excitation is due to glottal vibration. While excitation is present throughout the production process, it is considered significant (especially during glottal vibration) only when there is large energy in short-time interval, i.e., when it is impulse-like. These impulse-like characteristics are usually exhibited around the instants of glottal closure during each glottal cycle. The presence of these impulse-like characteristics suggests that the excitation can be approximated as a sequence of impulses. This assumption on the excitation of the vocal-tract system suggests a new approach for processing the speech signal as discussed in this section.

All physical systems are inertial in nature. The inertial systems respond when excited by an external source. The excitation to an inertial system can be any of the following four types.

1) *Excitation impulse is not in the observed interval of the signal—Sinusoidal generator:* Output signal is the response of a passive inertial system for an impulse, and the impulses themselves are not present in the observed intervals of the signal.
2) *Sinusoidal excitation:* Sinusoidal excitation can be viewed as impulse excitation in the frequency domain. Hence, a sinusoidal excitation to an inertial system selects the corresponding frequency component from transfer function of the system. Though sinusoidal excitation is widely used to analyze synthetic systems, it is not commonly found in physical systems.
3) *Random excitation:* Random excitation can be interpreted as impulse excitation of arbitrary amplitude at every instant of time. Since impulse excitations are present over all the instants of time, it is difficult to observe them from the output of the system. Random excitation does not possess impulse-like nature either in the time-domain or in the frequency-domain, and hence, the impulses cannot be perceived.
4) *Sequence of impulses as excitation:* In this case, the signals are generated by a passive inertial system with a fixed sequence of (periodic and/or aperiodic) impulses as excitation. The time instants of impulses may not be observed from the output of the system, but they can be perceived. If the sequence of impulses is periodic in the time-domain, then it corresponds to a periodic sequence of impulses in the frequency-domain also, and can also be perceived.

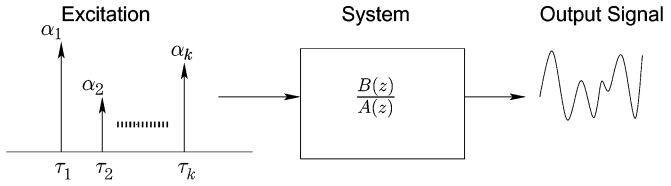

Fig. 1. Inertial system excited with a sequence of impulses.

Consider a physical system excited by a sequence of impulses of varying strengths, as shown in Fig. 1. One of the challenges in the field of signal processing is to detect the time instants (τ_k) of the impulses and their corresponding strengths (α_k) from the output signal. In a natural scenario like speech production, the characteristics of the system vary with time and are unknown. Hence, the signal processing problem can be viewed as a blind deconvolution, where neither the system response nor the excitation source are known. In this paper, we attempt to detect the time instants of excitation (epochs) of the vocal-tract system.

Consider a unit impulse in the time domain. It has all the frequencies equally well represented in the frequency domain. When an inertial system is excited by an impulse-like excitation, the effect of the excitation spreads uniformly in the frequency domain and is modulated by the time-varying transfer function of the system. The information about the time instants of occurrence of the excitation impulses reflects as discontinuities in the time domain. It may be difficult to observe these discontinuities directly from the signal because of the time-varying response of the system. The effect of the discontinuities can be highlighted by filtering the output signal through a narrowband filter centered around a frequency. The output of the narrowband filter predominantly contains a single frequency component, and as a result, the discontinuities due to the excitation impulses will get manifested as a deviation from the center frequency. The time instants of the discontinuities can be derived by computing the instantaneous frequency of the filtered output [24]. A tutorial review on the instantaneous frequency and its interpretation is given in [25]. It has been previously observed that isolated narrow spikes in the instantaneous frequency of the bandpass-filtered output [26, ch. 11] are attributed to either the valleys in the amplitude envelope or the onset of a new pitch pulse, but no previous work explored the feasibility of this type of observation for epoch extraction.

A. Computation of Instantaneous Frequency

The instantaneous frequency of a real signal $s(t)$ is defined as the time derivative of the unwrapped phase of the complex analytic signal derived from $s(t)$ [24]. The complex analytic signal corresponding to a real signal $s(t)$ is given by

$$s_a(t) = s(t) + j s_h(t) \tag{1}$$

where $s_h(t)$ is the Hilbert transform of the real signal $s(t)$ and is given by

$$s_h(t) = \mathrm{IFT}\left(S_h(\omega)\right) \tag{2}$$

where IFT denotes the inverse Fourier transform, and $S_h(\omega)$ is given by

$$S_h(\omega) = \begin{cases} +jS(\omega), & \omega < 0 \\ -jS(\omega), & \omega > 0. \end{cases} \tag{3}$$

The analytic signal thus derived contains only positive frequency components. The analytic signal $s_a(t)$ can be rewritten as

$$s_a(t) = |s_a(t)|\, e^{j\phi(t)} \tag{4}$$

where

$$|s_a(t)| = \sqrt{s^2(t) + s_h^2(t)} \tag{5}$$

is called the amplitude envelope, and

$$\phi(t) = \arctan\left(\frac{s_h(t)}{s(t)}\right) \tag{6}$$

is called the instantaneous phase. Direct computation of the phase $\phi(t)$ from (6) suffers from the problem of phase wrapping, i.e., $\phi(t)$ is constrained to an interval $(-\pi, \pi]$ or $[0, 2\pi)$. Hence, the instantaneous frequency cannot be computed by explicit differentiation of phase $\phi(t)$ without first performing the complex task of unwrapping the phase in time. The instantaneous frequency can be computed directly from the signal, without going through the process of phase unwrapping, by exploiting the Fourier transform relations. Taking logarithm on both sides of (4), and differentiating with respect to time t, we have

$$\begin{aligned} \log s_a(t) &= \log|s_a(t)| + j\phi(t) \\ \frac{s_a'(t)}{s_a(t)} &= \frac{d}{dt}\log|s_a(t)| + j\phi'(t) \end{aligned} \tag{7}$$

where the superscript $'$ denotes the derivative operator, and $\phi'(t)$ is the instantaneous frequency. That is

$$\phi'(t) = -\Im\left(\frac{s_a'(t)}{s_a(t)}\right) \tag{8}$$

where $\Im(.)$ denotes the imaginary part. $s_a'(t)$ can be computed by using the Fourier transform relations. The analytic signal $s_a(t)$ can be synthesized from its frequency domain representation through the inverse Fourier transform

$$s_a(t) = \frac{1}{2\pi}\int_0^\infty S_a(\omega)e^{j\omega t}d\omega \tag{9}$$

where $S_a(\omega)$ is the Fourier transform of the analytic signal $s_a(t)$, and is zero for negative frequencies. Differentiating both sides of (9) with respect to time t, we have

$$\begin{aligned} s_a'(t) &= \frac{1}{2\pi}\int_0^\infty S_a(\omega)e^{j\omega t}(j\omega)d\omega \\ &= j\left(\frac{1}{2\pi}\int_0^\infty (\omega S_a(\omega))\, e^{j\omega t}d\omega\right) \\ &= j\mathrm{IFT}\left(\omega S_a(\omega)\right). \end{aligned} \tag{10}$$

The instantaneous frequency $\phi'(t)$ can be obtained from (7) and (10) as

$$\phi'(t) = \Re\left(\frac{\mathrm{IFT}\left(\omega S_a(\omega)\right)}{\mathrm{IFT}\left(S_a(\omega)\right)}\right) \tag{11}$$

where $\Re(.)$ denotes real part. Computation of the instantaneous frequency given in (11) is implemented in the discrete domain as follows:

$$\phi'[n] = \frac{2\pi}{N}\Re\left(\frac{\mathrm{IDFT}\left(kS_a[k]\right)}{\mathrm{IDFT}\left(S_a[k]\right)}\right). \tag{12}$$

Here, IDFT denotes the inverse discrete Fourier transform, and N is the total number of samples in the signal.

The instantaneous frequency may be interpreted as the frequency of a sinusoid which locally fits the signal under analysis. However, it has a physical interpretation only for monocomponent signals, where there is only one frequency or a narrow range of frequencies varying as a function of time. In this case, the instantaneous frequency can be interpreted as deviation of frequency of the signal from the monotone at every instant of time. The notion of a single-valued instantaneous frequency becomes meaningless for multicomponent (multiple frequency sinusoids) signals. The multicomponent signal has to be dispersed into its components for further analysis.

In this paper, we propose to use a resonator to filter out from a signal a monocomponent centered around a single frequency for further analysis. A resonator is a second-order infinite-impulse response (IIR) filter with a complex conjugate pair of poles in the z-plane [27]. A resonator with narrow bandwidth (corresponding to a radius $r = 0.999$) was chosen to realize the narrow band filter. An ideal resonator $(r = 1)$ was not used in order to avoid saturation of the filter output.

When a multicomponent signal is filtered through a resonator centered around a frequency (ω_0), the output signal predominantly contains the ω_0 frequency component. Any deviation from ω_0 in frequency of the filtered output can be attributed to the impulse-like characteristics present in the multicomponent signal. In general, the analytic signal corresponding to the filtered output can be expressed as

$$s_a(t) = |s_a(t)|\, e^{j(\omega_0 t + \theta(t))}. \tag{13}$$

Hence, the instantaneous frequency of the filtered output (predominantly monocomponent) is given by

$$\phi'(t) = \omega_0 + \theta'(t). \tag{14}$$

Fig. 2(a) shows a multicomponent signal in the form of an aperiodic sequence of impulses with arbitrary strengths. The signal filtered by a 500-Hz resonator, and the instantaneous frequency of the filtered output are shown in Fig. 2(b) and (c), respectively. At the instants of impulse locations, the instantaneous frequency deviates significantly from the normalized center frequency $\omega_0 = 2\pi f/f_s$, where f is the frequency of the resonator, and f_s is the sampling frequency. For a resonator frequency $f = 500$ Hz, and sampling frequency $f_s = 8000$, the instantaneous frequency (around $\omega_0 = 0.39$) shows sharp peaks at the locations of the impulses. The illustration in Fig. 2 shows that the discontinuity information can be derived from the filtered output even if the sequence of impulses are not regularly spaced,

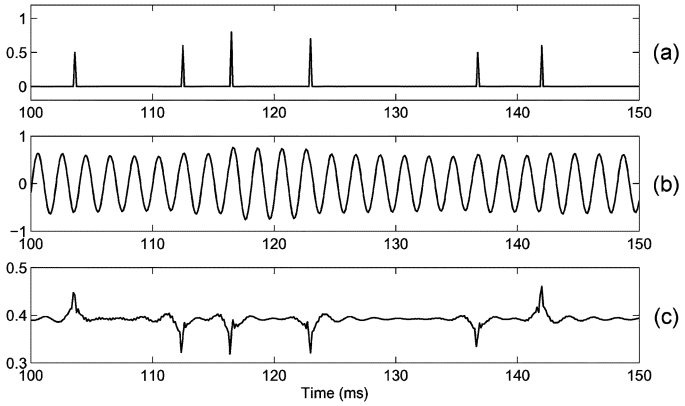

Fig. 2. Aperiodic sequence of impulses filtered through a 500-Hz resonator. (a) Aperiodic sequence of impulses with arbitrary strengths, (b) output of the resonator, and (c) instantaneous frequency of the filtered output.

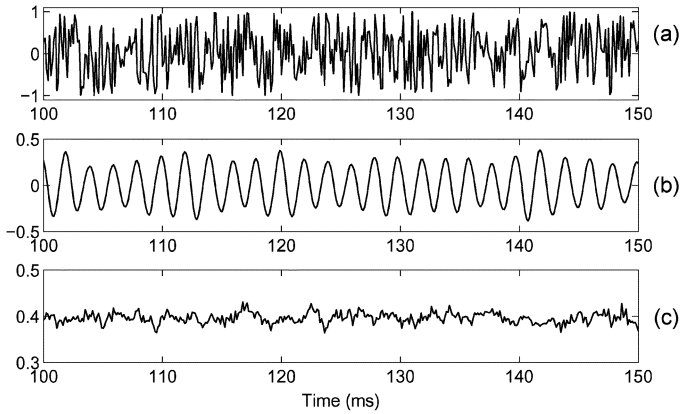

Fig. 3. White noise filtered through a 500-Hz resonator. (a) Segment of white noise, (b) output of the resonator, and (c) instantaneous frequency of the filtered output.

and are of arbitrary strengths. The amplitudes of the peaks in the instantaneous frequency depend not only on the strengths of the impulses, but also on the phases at which the sinusoids originated at these impulses are added at the instants. This in turn depends on the locations of the impulses and the frequency of the sinusoid.

To highlight the significance of these isolated discontinuities in the impulse sequence, if the impulse sequence is replaced by white noise, the corresponding filtered output and the IF plot do not contain any significant discontinuities, as shown in Fig. 3. The white noise does not contain any isolated impulse-like discontinuities. As a result, the filtered output will be a slowly varying amplitude envelope modulated by a sinusoid without any significant discontinuities in the phase. Hence, the instantaneous frequency of the filtered white noise does not show any significant peaks, unlike in the case of Fig. 2(c).

Consider a situation where a synthetic speech signal is filtered through a resonator. The synthetic speech signal is generated by exciting a time-varying all-pole system by a sequence of impulses at known locations. When such a signal is filtered through a resonator, the frequency response of the all-pole system gets multiplied with the frequency response of the resonator. Hence, the frequency response of the all-pole system around the center

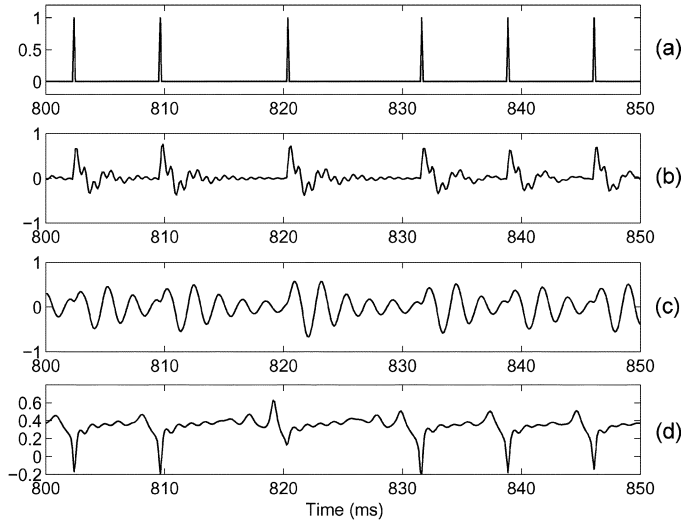

Fig. 4. Synthetic speech signal with known locations of excitation impulses filtered though a 500-Hz resonator. (a) Excitation impulses, (b) synthetic speech signal obtained by exciting an all-pole system with excitation impulses, (c) output of the resonator, and (d) instantaneous frequency of the filtered output.

frequency of the resonator gets selected. The filtered output carries the information about the discontinuities that are reflected in the narrow frequency band of the resonator. The instants of excitation impulses can be extracted from the filtered output using the instantaneous frequency. Fig. 4(b) shows a synthetic speech signal, obtained by exciting a time-varying all-pole system with a sequence of impulses shown in Fig. 4(a). The instantaneous frequency [Fig. 4(d)] of the filtered output [Fig. 4(c)] shows discontinuities at the instants of excitation of the all-pole system. The locations of the discontinuities are in close agreement with the original excitation impulses.

III. Illustration of Instantaneous Frequency for Speech Data

A speech signal can be considered as a convolution of the time-varying vocal-tract transfer function and the epochs due to the excitation source. The epochs are the time instants where significant excitation is delivered to the vocal-tract system. The information about the locations of the epochs is embedded in the coupling between the source and the system, though it is not evident from the speech waveform directly. It is difficult to accurately locate the time instants of excitation impulses directly from the speech waveform because of the time-varying resonances of the vocal-tract system. To highlight the effect due to the instants of significant excitation, the speech signal is filtered through a resonator centered around a chosen frequency ω_0. The significant deviations of the filtered output from the natural oscillations of the resonator can be attributed to the excitation impulses. Fig. 5 shows a 100-ms segment of voiced speech signal sampled at 8 kHz, and the output of the resonator at 500 Hz. The instantaneous frequency of the filtered output shows sharp peaks at the epoch locations, as shown in Fig. 5(c). In order to determine the accuracy of the estimated epoch locations, the differenced electro-glottograph (EGG) signal is also given in Fig. 5(d). The peaks in the instantaneous frequency of the filtered output match well with the actual epoch locations given

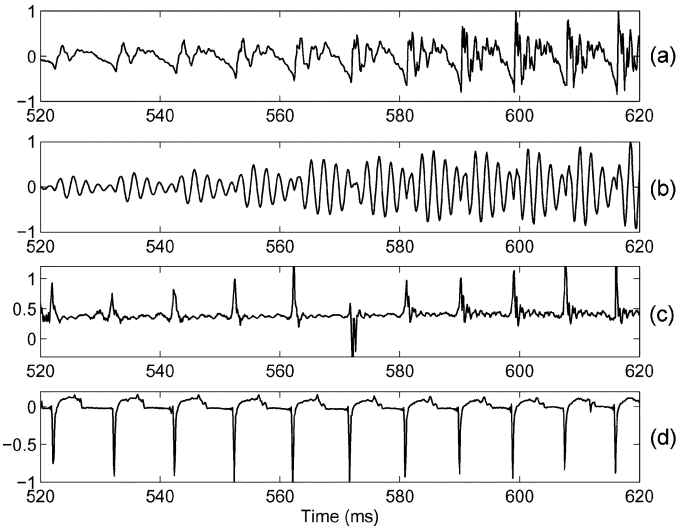

Fig. 5. A 100-ms segment of (a) speech waveform, (b) output of the resonator at 500 Hz, (c) instantaneous frequency of the filtered output, and (d) differenced EGG signal.

by the differenced EGG signal, illustrating the potential of the proposed method.

In the case of speech, instantaneous frequency of the filtered output also contains the time-varying frequency changes associated with the vocal-tract transfer function, which is undesirable. As a result, though the peaks in the instantaneous frequency of the filtered output indicate the epoch locations accurately for the segment shown in Fig. 5, it may not be useful to extract the epoch locations unambiguously for any chosen center frequency (ω_0). Thus, the method of epoch extraction using the instantaneous frequency of the filtered output depends critically on the choice of center frequency of the filter. A single center frequency may not be suitable for extracting the epoch locations of an arbitrary segment of speech. The center frequency has to be chosen based on the characteristics of the speech segment under analysis. The choice of the center frequency also depends on the distribution of energy of the speech segment in the frequency domain. To illustrate the significance of choice of the center frequency of the filter, the instantaneous frequency computed around four different center frequencies are shown in Fig. 6. The spectrogram, the speech signal and the differenced EGG signal are also given for reference. The spectrogram in Fig. 6(a) shows a band of energy around 500 Hz. The instantaneous frequency computed around 500 Hz [Fig. 6(d)] indicates unambiguous peaks/valleys that are in close agreement with the actual epochs shown by the differenced EGG signal [Fig. 6(c)]. In the instantaneous frequencies computed around 1000 and 2000 Hz, shown in Fig. 6(e) and (f), respectively, the epoch locations cannot be identified easily. This is because the energy of the signal in those frequency bands is very low. Since the spectrogram shows large energy in the band around 2500 Hz, the instantaneous frequency computed around 2500 Hz shows sharp peaks/valleys around the epoch locations. However, the instantaneous frequency plot in Fig. 6(g) shows less ambiguous peaks/valleys in the time interval 570–620 ms, than those in the time interval 520–570 ms. This is because the intensity of the 2500-Hz frequency band in the time interval 570–620 ms is greater than the intensity of the band in the time interval 520–570 ms.

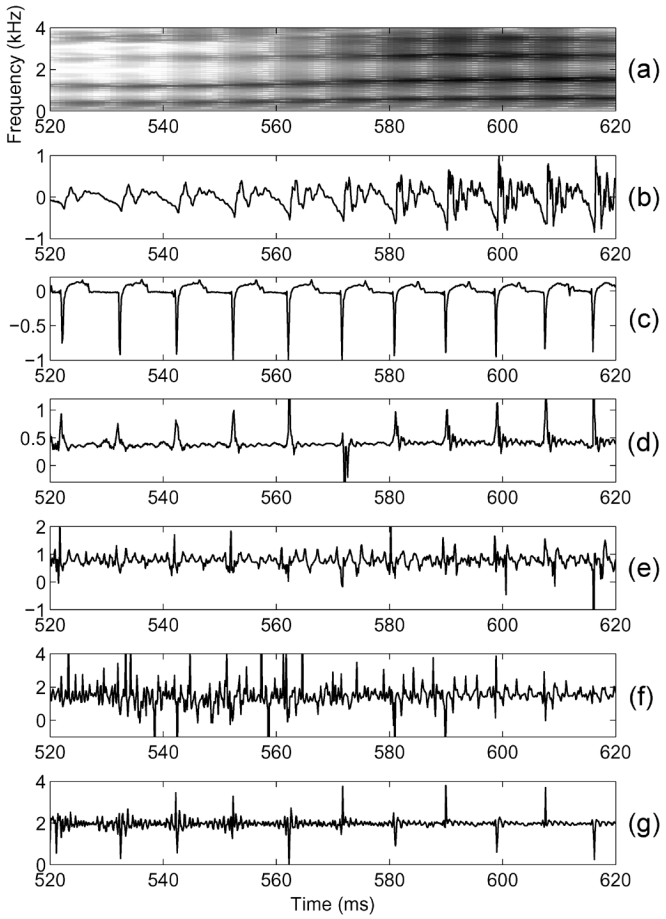

Fig. 6. Instantaneous frequency of a speech segment computed around four different center frequencies. (a) Spectrogram of the speech segment. (b) Speech waveform. (c) Differenced EGG signal. Instantaneous frequency plots computed around (d) 500 Hz, (e) 1000 Hz, (f) 2000 Hz, and (g) 2500 Hz.

Notice that the instantaneous frequencies computed around 1000 and 2000 Hz also contain all the peaks/valleys corresponding to the epoch locations, but they cannot be located easily due to fluctuations in the neighborhood. This is because the instantaneous frequency captures not only the discontinuities due to the excitation impulses, but also the fluctuations due to the time-varying vocal-tract system. Hence, it is difficult to extract the instants of excitation from the instantaneous frequency computed around an arbitrary center frequency. The center frequency has to be chosen in such a way that the discontinuities due to the excitation impulses dominate over the fluctuations due to the time-varying vocal-tract system.

IV. Epoch Extraction From Speech Using a 0-Hz Resonator

The discontinuity due to impulse excitation is reflected across all the frequencies including the zero frequency. That is, even the output of the resonator at zero frequency should have the information of the discontinuities due to impulse-like excitation. The advantage of choosing the zero frequency resonator filter is that the characteristics of the time-varying vocal-tract system will not affect the characteristics of the discontinuities in the resonator filter output. This is because the vocal-tract system has resonances at much higher frequencies than at zero frequency.

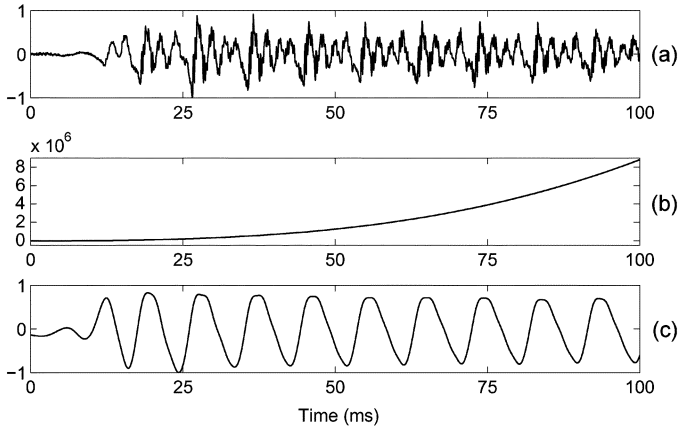

Fig. 7. Illustration of effect of mean subtraction from output of 0-Hz resonator. A 100-ms segment of (a) speech signal, (b) output of cascade of two 0-Hz resonators, and (c) mean subtracted signal.

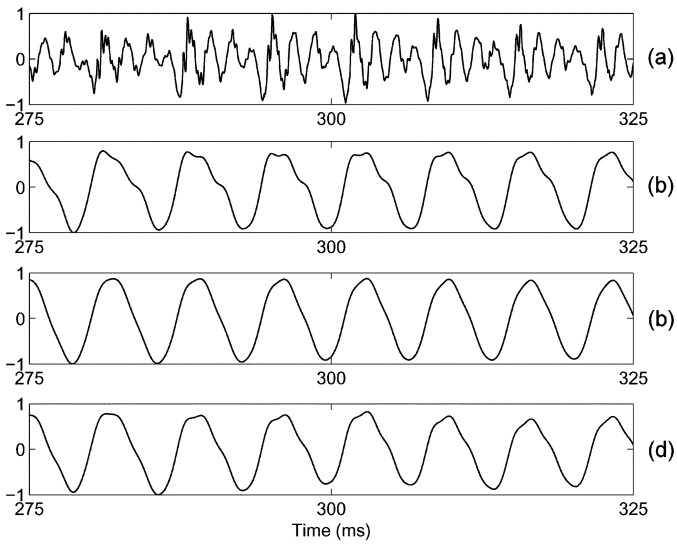

Fig. 8. Illustration of effect of length of window used for mean subtraction. (a) Speech signal. Mean subtracted signal using a window length of (b) 5 ms, (c) 10 ms, and (d) 15 ms.

Therefore, we propose that the characteristics of the discontinuities due to excitation impulses can be extracted by passing the speech signal twice through a zero frequency filter. The purpose of passing the speech signal twice is to reduce the effects of all (high frequency) resonances. A cascade of two 0-Hz resonators provide a sharper roll-off compared to a single 0-Hz resonator. Since the output of the zero frequency filter is equivalent to double integration of the signal, passing the speech signal twice through the filter is equivalent to four times successive integration. This will result in a filtered output that grows/decays as a polynomial function of time. Fig. 7 shows a segment of speech signal, and its filtered output. The effect of discontinuities due to impulse sequences will be overridden by those large values of the filtered output. Hence, it is difficult to compute the instantaneous frequency (deviation from zero frequency) as in the conventional manner of computing the analytic signal of the filtered output.

We attempt to compute the deviation of the filtered output from the local mean to extract the characteristics of the discontinuities due to impulse excitation. The local mean for every 10 ms is computed and is subtracted from the filtered output. The resulting mean subtracted signal obtained from the filtered output in Fig. 7(b) is shown in Fig. 7(c). The mean subtracted signal is called 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.

1) Difference the speech signal $s[n]$ (to remove any time-varying low frequency bias in the signal)

$$x[n] = s[n] - s[n-1] \tag{15}$$

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

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{16a}$$

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \tag{16b}$$

where $a_1 = -2$, and $a_2 = 1$. This is equivalent to successive integration four times, but we prefer to call the process as filtering at zero frequency.

3) Remove the trend in $y_2[n]$ by subtracting the average over 10 ms at each sample. The resulting signal

$$y[n] = y_2[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_2[n+m] \tag{17}$$

is called the zero-frequency filtered signal, or simply the filtered signal. Here $2N+1$ corresponds to the number of samples in the 10 ms interval.

The effect of the time window for local mean computation is shown in Fig. 8 for 5, 10, and 15 ms. The choice of the window size is not critical in the range of 5-15 ms. It is preferable to have a window size of one to two pitch periods to avoid spurious zero crossings in the filtered signal.

The filtered signal clearly shows rapid changes around the positive zero crossings. So the time instants of the positive zero crossings can be used as epochs. It is interesting to note that for impulse sequences (even for aperiodic sequences) the positive zero-crossing instants correspond to the locations of the impulses. There is no such relation between the excitation and the filtered signal for the random noise excitation of the time-varying all-pole system. Also, the filtered signal has significantly lower values for the random noise excitation compared to the impulse sequence excitation. Fig. 9(b) shows the filtered signal for a speech signal consisting of voiced and unvoiced segments. The unvoiced segments correspond to the random noise excitation of the vocal-tract system. The differenced EGG signal [Fig. 9(c)] is also given in the figure to identify the voiced and nonvoiced segments. Fig. 10 shows the speech waveform, the filtered signal and the derived epoch locations and the differenced EGG signals for an utterance of a female voice. The epoch locations coincide with the locations of the large negative peaks in the differenced EGG signal [Fig. 10(c)].

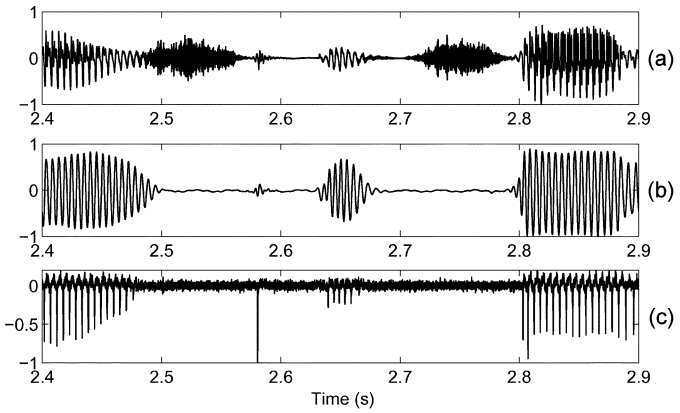

Fig. 9. Segment of (a) speech signal, (b) filtered signal, and (c) differenced EGG signal. The filtered output shows significantly lower values in the regions where there is no glottal activity.

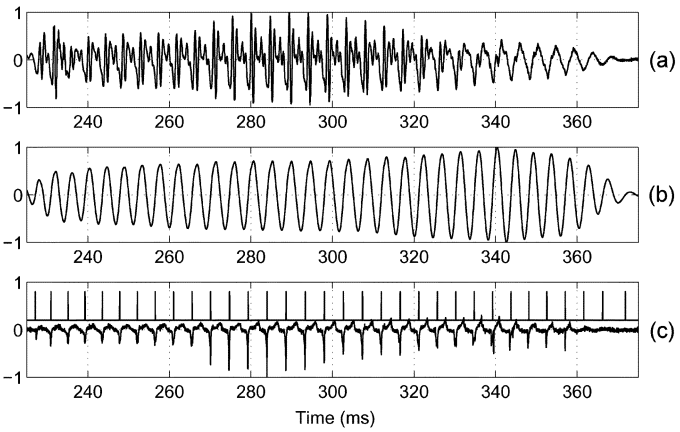

Fig. 10. Illustration of the proposed method of epoch detection for female speaker. (a) Speech signal, (b) filtered signal, and (c) differenced EGG signal. Pulses in (c) indicate the detected epochs. Note that the filtered output brings out even the epochs not picked up by the EGG signal (in the interval 360–375 ms).

V. Comparison of Proposed Epoch Extraction With Other Methods

In this section, the proposed method of epoch extraction is compared with three existing methods in terms of identification accuracy and in terms of robustness against degradation. The three methods chosen for comparison are the Hilbert envelope-based (HE-based) method [28], the group-delay-based (GD-based) method [18], and the DYPSA algorithm [23]. Initially, the performance of the algorithms was evaluated on the clean data. Subsequently, we have evaluated robustness of the proposed method and the three existing methods at different levels of degradations. A brief discussion on the implementation details of the three chosen methods for comparison are given below.

Hilbert envelope-based method: The strength of the excitation impulses in the voiced speech is large and impulse-like. Though this can be observed from the LP residual, it cannot be extracted unambiguously because of multiple peaks of random polarity around the instant of excitation. Ideally, it is desirable to derive an impulse-like signal around the instant of significant excitation. A close approximation to this is possible by using the Hilbert envelope of the LP residual. Even though the real and imaginary parts of an analytic signal have positive and negative

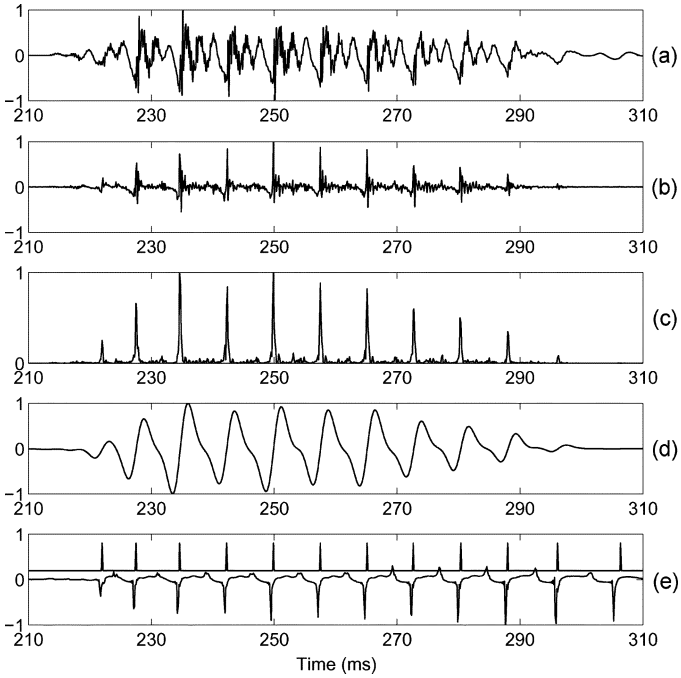

Fig. 11. Illustration of Hilbert envelope-based method for epoch extraction [28]. (a) Speech signal, (b) LP residual, (c) Hilbert envelope of LP residual, (d) epoch evidence plot, and (e) differenced EGG signal. The pulses in (e) indicate the detected epoch locations.

samples, the Hilbert envelope of a signal is a positive function, giving the envelope of the signal. For example, the HE of a unit sample sequence or its derivative has a peak at the same instant. Thus, the properties of the HE can be exploited to derive approximate epoch locations. The evidence for epoch locations can be obtained by convolving the HE with a Gabor filter (modulated Gaussian pulse), as suggested in [28]. In the present work, the evidence for epoch locations is obtained by convolving the HE with a differenced Gaussian pulse

$$g[n] = \frac{(n - N/2)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(n - N/2)^2}{2\sigma^2}\right),\ n = 1, 2, \ldots, N,$$

where σ defines the spatial spread of the Gaussian, and N is the length of the filter. For this evaluation, the values of $\sigma = 10$, and $N = 10$ ms (80 samples at 8-kHz sampling frequency) are used. The Hilbert envelope of the LP residual is convolved with the differenced Gaussian pulse to obtain the epoch evidence plot shown in Fig. 11(d). The instants of positive zero crossings in the epoch evidence plot correspond approximately to the locations of the instants of significant excitation.

Group delay-based method: This method is based on the global phase characteristics of minimum phase signals. The average slope of the unwrapped phase of the short-time Fourier transform of LP residual is computed as a function of time. The averaged slope obtained as a function of time is termed as phase-slope function. Instants where the phase-slope function makes a positive zero crossing are identified as epochs. Fig. 12 shows a speech utterance, its LP residual, the phase-slope function, and the extracted instants. For this evaluation, we have used a tenth-order LP analysis to derive the LP residual, and an 8-ms window for computing the phase-slope function.

The DYPSA algorithm: The DYPSA algorithm is an automatic technique for estimating the epochs in voiced speech

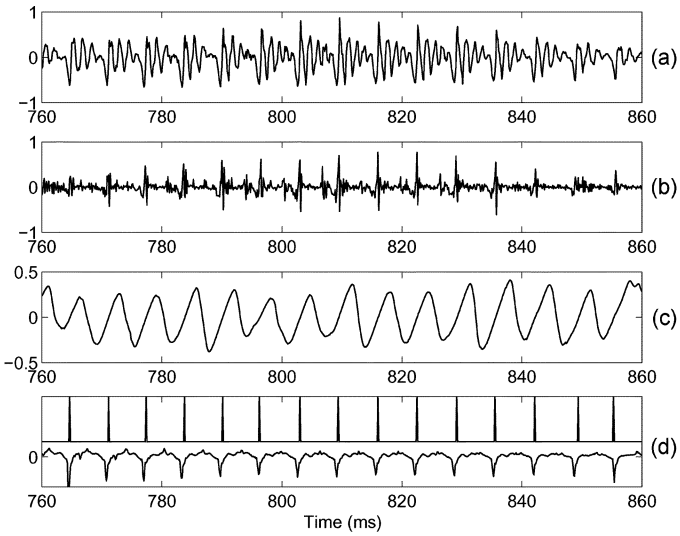

Fig. 12. Illustration of group-delay based method for epoch extraction [18]. (a) Speech signal, (b) LP residual, (c) phase-slope function, and (d) differenced EGG signal. The pulses in (d) indicate the detected epoch locations.

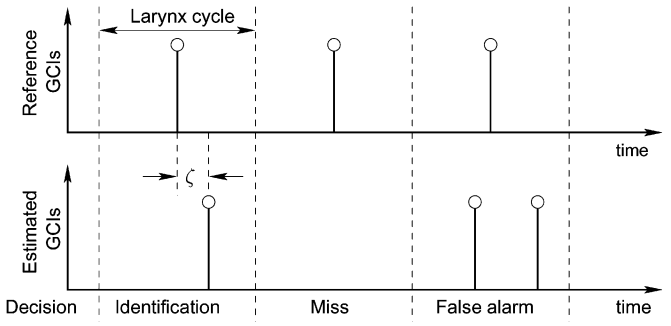

Fig. 13. Characterization of epoch estimates showing three larynx cycles with examples of each possible outcome from epoch extraction [23]. Identification accuracy is measured as a variance of ζ.

TABLE I
PERFORMANCE COMPARISON OF EPOCH DETECTION METHODS ON CMU-ARCTIC DATABASE. IDR—IDENTIFICATION RATE, MR—MISS RATE, FAR—FALSE ALARM RATE, IDA—IDENTIFICATION ACCURACY

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
HE-based	89.86	1.43	8.71	0.58
GD-based	92.8	4.01	3.18	0.67
DYPSA	96.66	1.76	1.58	0.59
Proposed	99.04	0.18	0.77	0.36

from the speech signal alone. There are three components in the algorithm. The first component generates candidate epochs using zero crossings of the phase-slope function. The energy weighted group-delay was used as a measure to derive the phase-slope function. The second component employs a novel phase-slope projection technique to recover candidates for which the phase-slope function does not include a zero-crossing. These two components detect almost all the true epochs, but they also generate a large number of false alarms. The third component of the algorithm uses dynamic programming to identify the true epochs from the set of hypothesized candidates by minimizing a cost function. For evaluating this technique, the MATLAB implementation of the DYPSA available in [29] was used.

The CMU-Arctic database [30], [31] was employed to evaluate the proposed method of epoch detection and to compare the results with the existing methods. The Arctic database consists of 1132 phonetically balanced English sentences spoken by two male and one female talkers. The duration of each utterance is approximately 3 s, which makes the duration of the entire database to be around 2 h 40 min. The database was collected in a soundproof booth, and digitized at a sampling frequency of 32 kHz. In addition to the speech signals, the Arctic database contains the simultaneous recordings of EGG signals collected using a laryngograph. The speech and EGG signals were time-aligned to compensate for the larynx-to-microphone delay, determined to be approximately 0.7 ms. Reference locations of the epochs were extracted from the voiced segments of the EGG signals by finding peaks in the differenced EGG signal. The performance of the algorithms was evaluated only in the voiced segments (detected from EGG signal) between the reference epoch locations and the estimated epoch locations. The database contains a total of 792 249 epochs in the voiced regions.

The performance of the epoch detection methods was evaluated using the measures defined in [23]. Fig. 13 shows the characterization of epoch estimates showing each of the possible decisions from the epoch detection algorithms. The following measures were defined to evaluate the performance of epoch detection algorithms.

1) *Larynx cycle:* The range of samples $(1/2)(l_{r-1}+l_r) \leq n \leq (1/2)(l_r + l_{r+1})$, given an epoch reference at sample l_r with preceding and succeeding epoch references at samples l_{r-1} and l_{r+1}, respectively.
2) *Identification rate* (IDR): The percentage of larynx cycles for which exactly one epoch is detected.
3) *Miss rate* (MR): The percentage of larynx cycles for which no epoch is detected.
4) *False alarm rate* (FAR): The percentage of larynx cycles for which more than one epoch is detected.
5) *Identification error* ζ: The timing error between the reference epoch location and the detected epoch location in larynx cycles for which exactly one epoch was detected.
6) *Identification accuracy* σ (IDA): The standard deviation of the identification error ζ. Small values of σ indicate high accuracy of identification.

Table I shows the performance results on Arctic database for identification rate, miss rate, false alarm rate, and identification accuracy for the three methods HE-based, GD-based, and DYPSA algorithm, as well as for the proposed method. Fig. 14 shows the histograms of the timing errors ζ in detecting the epoch locations, averaged over the entire Arctic database. From Table I, it can be concluded that the DYPSA algorithm performed best among the three existing techniques, with an identification rate of 96.66%. The proposed method of epoch detection gives a better identification rate as well as identification accuracy, compared to the results from the DYPSA algorithm.

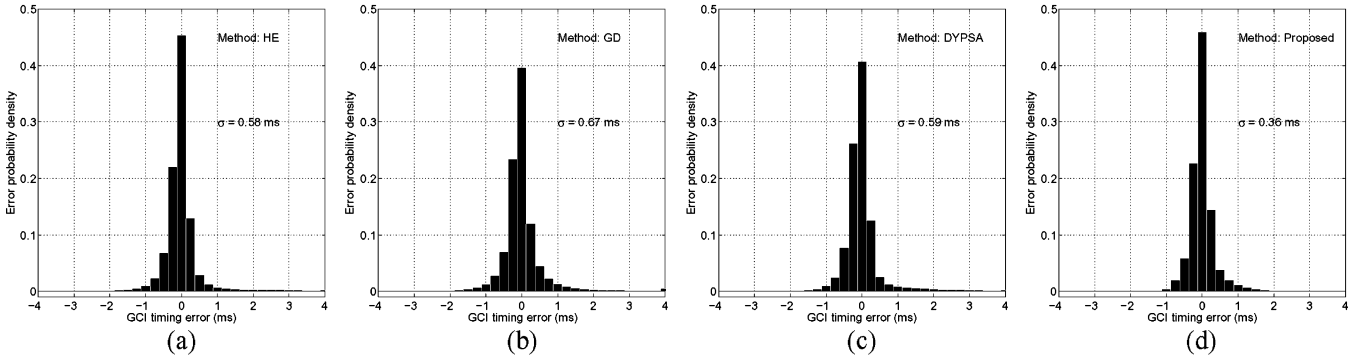

Fig. 14. Histogram of the epoch timing errors for (a) HE-based method, (b) GD-based method, (c) DYPSA algorithm, and (d) proposed method.

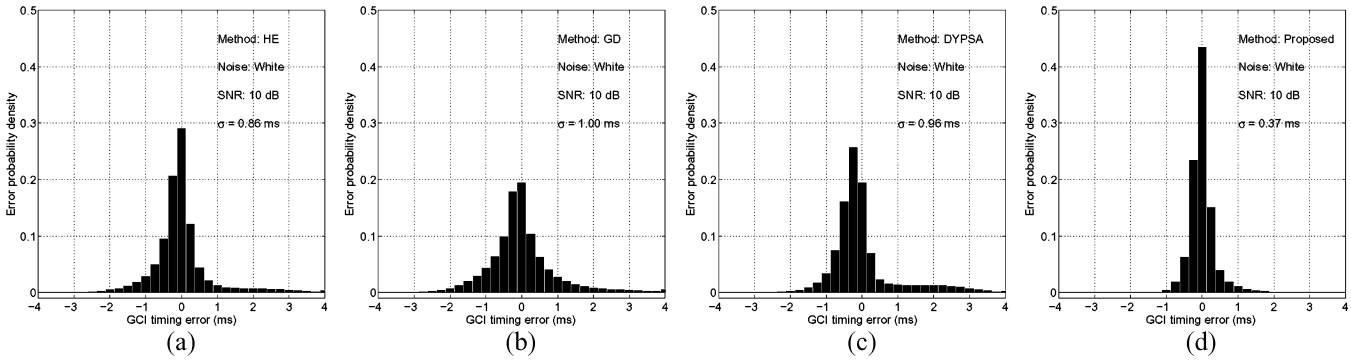

Fig. 15. Histogram of the epoch timing errors for degradation by white noise at an SNR of 10 dB. (a) HE-based method, (b) GD-based method, (c) DYPSA algorithm, and (d) proposed method.

VI. Effect of Noise on Performance of Proposed Method of Epoch Extraction

In this section, we study the effect of (moderate levels of) noise on the accuracy of the epoch detection methods. The existing methods and the proposed method are evaluated on an artificially generated noisy speech database. Several noise environments at varying signal-to-noise ratio (SNR) were simulated to evaluate the epoch detection methods. The noise used was taken for NOISEX-92 database [32]. The database consists of white, babble, high-frequency (HF) channel, and vehicle noise. The noise from the NOISEX-92 database was added to the Arctic database to form noisy speech data at different levels of degradation. The utterances are appended with silence such that the total amount of silence in each utterance is constrained to be about 60% of data, including the pauses in the utterances. Including different noise environments and SNRs, the database consists of 33 h of noisy speech data.

Table II shows the comparative results of epoch detection methods for different types of degradations at varying SNRs. Fig. 15 shows the distribution of the timing errors ζ in detecting the epoch locations, for white noise environment at an of SNR of 10 dB. The proposed method consistently performs better than the existing techniques even under degradation. The improved performance of the proposed method may be attributed to the following reasons. 1) There is no block processing involved in this method. Hence, there are no effects of the size and the shape of the window. The entire speech signal is processed at once to obtain the filtered signal. 2) The proposed method is not dependent on the energy of the signal. This method detects the epoch locations even in weakly voiced regions like voice-bar. 3) There is only one parameter involved in the proposed method, i.e., the length of the window for removing the trend from the output of 0-Hz resonator. 4) There are no critical thresholds or costs involved in identifying the epoch locations.

VII. Summary and Conclusion

In this paper, we proposed a method for epoch extraction that does not depend on the characteristics of the vocal-tract system. The method exploits the impulse-like excitation of the vocal-tract system. The method uses the output of speech from a zero frequency resonator. The positive zero crossings of the filtered signal correspond to epochs. The identification rate and identification accuracy are evaluated using the CMU-Arctic database, where the speech signal and the corresponding EGG signals are available. The epoch information derived from the EGG signals is used as a reference. The performance of the proposed method is compared with the results from the DYPSA and two other methods. The proposed method gives significantly better results in terms of identification rate and identification accuracy. It is also interesting to note that the proposed method is robust against degradations such as white noise, babble, high-frequency channel, and vehicle noise.

There are many novel features in the proposed method of epoch extraction. The method does not use any block processing as most signal processing methods do. The performance of the method does not depend on the energy of the segment of speech signal, and hence, the method works equally well for all types of speech sound units. In addition, there are no parameters to control, and no arbitrary thresholding in the identification of epochs.

TABLE II
PERFORMANCE COMPARISON FOR EPOCH DETECTION METHODS FOR VARIOUS SNRS AND NOISE ENVIRONMENTS. IDR—IDENTIFICATION RATE, MR—MISS RATE, FAR—FALSE ALARM RATE, IDA—IDENTIFICATION ACCURACY

Environment		HE Based				GD Based			
Noise	SNR (dB)	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDR (%)	MR (%)	FAR (%)	IDA (ms)
White	20 dB	84.56	1.58	13.86	0.686	87.34	3.82	8.85	0.812
White	15 dB	82.26	1.9	15.85	0.761	84.65	4.15	11.2	0.891
White	10 dB	79.45	2.39	18.16	0.864	81.07	4.79	14.14	0.907
Babble	20 dB	86.73	1.54	11.73	0.674	89.45	3.99	6.56	0.782
Babble	15 dB	84.88	1.77	13.35	0.743	87.27	4.28	8.45	0.855
Babble	10 dB	82.51	2.17	15.32	0.842	84.32	4.77	10.91	0.956
HF Channel	20 dB	84.23	1.87	13.91	0.738	86.54	4.36	9.10	0.849
HF Channel	15 dB	82.04	2.26	15.69	0.822	83.87	4.84	11.29	0.934
HF Channel	10 dB	79.24	2.85	17.91	0.927	80.13	5.53	14.34	1.040
Vehicle	20 dB	89.75	1.40	8.85	0.584	92.67	3.95	3.38	0.674
Vehicle	15 dB	89.58	1.39	9.03	0.585	92.49	3.92	3.59	0.679
Vehicle	10 dB	89.25	1.37	9.38	0.591	92.18	3.88	3.95	0.689

Environment		DYPSA				Proposed Method			
Noise	SNR (dB)	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDR (%)	MR (%)	FAR (%)	IDA (ms)
White	20 dB	92.12	1.41	6.47	0.738	99.04	0.19	0.77	0.363
White	15 dB	85.33	1.24	13.43	0.841	99.06	0.19	0.75	0.365
White	10 dB	75.95	1.09	22.96	0.957	99.05	0.23	0.72	0.371
Babble	20 dB	96.42	1.8	1.79	0.621	99.02	0.19	0.79	0.366
Babble	15 dB	96.14	1.82	2.05	0.647	98.99	0.21	0.80	0.374
Babble	10 dB	95.48	1.78	2.74	0.69	98.83	0.30	0.87	0.405
HF Channel	20 dB	95.89	1.77	2.33	0.654	99.04	0.19	0.77	0.363
HF Channel	15 dB	94.99	1.66	3.35	0.702	99.05	0.19	0.76	0.363
HF Channel	10 dB	92.4	1.56	6.01	0.775	99.06	0.21	0.73	0.368
Vehicle	20 dB	96.67	1.76	1.57	0.589	99.06	0.20	0.73	0.372
Vehicle	15 dB	96.6	1.78	1.62	0.596	98.93	0.37	0.70	0.397
Vehicle	10 dB	96.64	1.76	1.61	0.597	97.83	1.53	0.64	0.460

The method performs well for speech collected with a close-speaking microphone, even with the addition of degradations. However, the method is not likely to work well when the degradations produce additional impulse-like sequences in the collected speech data as in the case of reverberation. The method is also not likely to work well when there is interference of speech from other speakers. Our future efforts will be in the direction of developing methods for extracting epochs from speech with degradations involving superposed impulse-like characteristics.

Since the proposed method provides accurate locations of epochs, the results are useful to develop methods for pitch extraction, and also for voice activity detection. Also, since the filtered signal gives an indication of glottal activity, the method may be used for analysis of phonation characteristics [33] in normal and pathological voices. The method may also be a useful first step in accurate analysis of vocal-tract characteristics by focusing the attention in the region around the epochs. Accurate analysis of excitation source and time-varying vocal-tract systems may lead to a better acoustic–phonetic analysis of speech sounds in many languages, and it also may provide a useful supplement to the existing spectral-based methods of speech analysis.

REFERENCES

[1] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Signal Process.*, vol. SP-33, no. 4, pp. 369–377, Apr. 1985.

[2] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, Jul. 1998.

[3] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time domain prosodic modifications of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Glasgow, U.K., May 1989, pp. 238–241.

[4] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.

[5] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswamy, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1110–1118, Nov. 2005.

[6] B. Yegnanarayana and P. S. Murty, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[7] A. Neocleous and P. A. Naylor, "Voice source parameters for speaker verification," in *Proc. Eur. Signal Process. Conf.*, 1998, pp. 697–700.

[8] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–56, Jan. 2006.

[9] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. New York : Springer-Verlag, 1982.

[10] A. N. Sobakin, "Digital computer determination of formant parameters of the vocal tract from a speech signal," *Soviet Phys.-Acoust.*, vol. 18, pp. 84–90, 1972.

[11] H. W. Strube, "Determination of the instant of glottal closures from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1625–1629, 1974.

[12] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 6, pp. 562–570, Dec. 1975.

[13] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637–655, 1971.

[14] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 309–319, Aug. 1979.

[15] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal closure inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, Aug. 1979.

[16] Y. M. Cheng and O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 12, pp. 1805–1815, Dec. 1989.

[17] Y. K. C. Ma and L. F. Willems, "A Fribenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 258–265, Apr. 1994.

[18] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[19] B. Yegnanarayana and R. L. H. M. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, May 1995, pp. 776–779.

[20] P. S. Murty and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant excitation from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 609–619, Nov. 1999.

[21] M. Brookes, P. A. Naylor, and J. Gundnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 2, pp. 456–466, Mar. 2006.

[22] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, May 2002, vol. 11, pp. 349–352.

[23] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[24] L. Cohen, *Time-Frequency Analysis: Theory and Applications*, ser. Signal Processing Series. Englewood Cliffs: Prentice-Hall, 1995.

[25] B. Boushash, "Estimating and interpreting the instantaneous frequency of a signal—Part 1: Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, Apr. 1992.

[26] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Singapore: Pearson, 2004.

[27] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[28] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group-delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, Oct. 2007.

[29] M. Brookes, Voicebox: A Speech Processing Toolbox for MATLAB. 2006. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[30] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224.

[31] "CMU-ARCTIC Speech Synthesis Databases." [Online]. Available: http://festvox.org/cmu_arctic/index.html

[32] "Noisex-92," [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[33] 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. Workshop Speech Anal. Process. Knowledge Discovery*, Aalborg, Denmark, Jun. 2008, Aalborg Univ..

K. Sri Rama Murty received the B.Tech 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.

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 at the 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 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.