J. Intell. Syst. **21** (2012), 79–95
DOI 10.1515/jisys-2012-0003

Time-Order Representation Based Method for Epoch Detection from Speech Signals

Pooja Jain and Ram Bilas Pachori

Abstract. Epochs present in the voiced speech are defined as time instants of significant excitation of the vocal tract system during the production of speech. Nonstationary nature of excitation source and vocal tract system makes accurate identification of epochs a difficult task. Most of the existing methods for epoch detection require prior knowledge of voiced regions and a rough estimation of pitch frequency. In this paper, we propose a novel method that relies on time-order representation (TOR) based on short-time Fourier–Bessel (FB) series expansion which can be employed on entire speech signal to detect epochs without any prior information. The proposed method automatically detects voiced regions in the speech signal by computing the marginal energy density with respect to time in the low frequency range (LFR) from the energy distribution in the time-frequency plane. An estimate of pitch frequency for each detected voiced region is then obtained by computing the marginal energy density with respect to frequency in the LFR from the energy distribution in the time-frequency plane. Epochs are located for each detected voiced region as peaks in the derivative of the low pass filtered (LPF) signal corresponding to falling edges of peak negative cycles in the LPF signal synthesized from TOR coefficients corresponding to LFR. Experimental results obtained by the proposed method on speech signals taken from the CMU-Arctic database are found to be promising. The proposed method detects epochs with high accuracy and reliability.

Keywords. Speech Signal Analysis, Epoch Detection, Pitch Frequency Estimation, Voiced Detection, Time-Order Representation, Fourier–Bessel Series Expansion.

2010 Mathematics Subject Classification. 42Cxx.

1 Introduction

Speech is produced when the vocal tract system is stimulated by the excitation source. Air is forced through lungs and passed through vocal folds, acting as an excitation source. The vocal tract system comprises of lips, teeth, nasal cavity, tongue and mouth which act as a time-varying filter amplifying certain frequencies and attenuating other frequencies present in the excitation. The speech signal consists of two parts, voiced and non-voiced. During voiced speech, the vocal folds vibrate and excitation takes the form of quasi-periodic puffs of air which produces output speech that appears periodic. Non-voiced speech consists of unvoiced speech and silence. Unvoiced speech is produced when air rapidly passes through a narrow constriction in the windpipe, resulting in white noise like output. Silence contains only background noise. About one third part of a speech signal is unvoiced [9]. Epochs are defined as time instants of significant excitation of the vocal tract system during the production of speech. Vocal folds vibrate during voiced activity and glottal closure causes sudden decrease in the glottal impedance resulting in high signal strength. Time instants of glottal closure when there is little or no air flow through the glottis are known as epochs or glottal closure instants (GCIs). Epochs can be detected from the electro-glottograph (EGG) signal but the EGG signal is normally not available in practical applications. This provides a strong motivation to develop techniques for detection of epochs directly from the speech signal. In this paper, we will use the word epoch and the acronym GCI interchangeably to denote instants of glottal closure.

Accurate epoch detection is a fundamental requirement for many speech analysis and synthesis applications. Instantaneous pitch frequency computation from GCIs can be used for automatic gender identification and emotion recognition [10, 14, 37]. Pitch synchronous waveform encoding of voice needs epoch information [12]. Precise identification of epochs allows blind deconvolution of vocal tract system and excitation source [36, 39]. Speaker recognition and diagnosis of voice disorders can be performed with knowledge of closed glottis regions [17,29]. The detected epochs can be employed as pitch markers for prosody manipulation, which is useful in applications like text to speech synthesis and voice conversion [8, 30].

Several methods have been developed so far to determine epochs from speech signal without the availability of EGG signal. The task of detecting epochs from speech signal was earlier addressed by Sobakin and Strube [33, 34]. Time instant corresponding to the maximum of the determinant of the auto-covariance matrix within a pitch period of the speech signal is recognized as an epoch. However, the method has high computational complexity, low accuracy and it does not work well for some vowel sounds. Many classical methods for GCI detection relied on a linear prediction model of speech signal, where the vocal tract system is modeled as an all-pole filter. The parameters of the filter are assumed to be stationary for the duration of 20–30 ms, assuming that the vocal tract system is varying slowly. Linear prediction (LP) residual signal is obtained by removing the linear prediction of the speech signal from the speech signal. Time instant corresponding to large value of LP residual within a pitch period is indicative of GCI [2] but the presence of peaks of opposite polarities in the LP residual signal around a GCI causes ambiguity. This limitation was overcome in [1] by computing the Hilbert envelope of the LP residual whose positive peaks undoubtedly locate GCIs. The drawback of LP based analysis is that it is sensitive to noise. Moreover, accurate parameter estimation of an all-pole model and characterization of the excitation source are interdependent problems.

Cheng et al. proposed a maximum likelihood theory based method for GCI detection [4]. In this method, the strongest positive pulse within a pitch period of the resultant signal is detected as GCI. Many suboptimal pulses are present in the close vicinity of the GCI candidate; therefore a selection function is derived from the speech signal and its Hilbert transform to contrast between GCI candidate and other suboptimal pulses. GCI locations from the speech signal can be extracted from its short-term energy estimate. The Frobenius norm based approach for GCI identification was proposed in [16]. The Frobenius norm computed using a sliding window provides an estimate of signal energy at each sample instant. Time instant with maximum energy within a pitch period is identified as GCI. However, simulations were carried out only on vowel segments; more complicated cases like semivowels, nasals have not been dealt with. Epochs can also be detected using time-frequency analysis. Wavelet transform obtained at various scales has been used to detect GCIs [35]. Amplitude maxima at various scales are organized into lines of maximal amplitude (LOMA) using a dynamic programming algorithm. These lines form “trees” in the time-scale domain. GCIs are then interpreted as the top of the strongest branch, or trunk, of these trees. Cohen’s class time-frequency representation based approach for detecting GCIs from speech signal in noisy environments was presented in [19]. A detection function has been defined and morphologic filtering has been applied over it to determine GCIs optimally.

Group delay measures have been used to identify GCIs [32]. Average group delay of the speech signal within an analysis frame corresponds to GCI. The resilience of group delay measure in noisy environment was studied in [38]. A quantitative assessment of various group delay measures for GCI identification from voiced speech signal was made in [3]. Different group delay measures, namely average group delay, zero frequency group delay and energy weighted group delay, were compared on the basis of various performance parameters such as computational complexity, detection rate, and accuracy. A dynamic programming projected phase-slope algorithm (DYPSA) for GCI detection from voiced speech signal was presented in [15, 20]. Zero crossings of the phase-slope function derived from the energy weighted group-delay are further refined by a dynamic programming algorithm and identified as GCIs. Lately methods that do not depend critically on characteristics of time-varying vocal tract system have been proposed. The zero frequency resonator (ZFR) was proposed in [18] to extract epochs from voiced speech signal. This method requires the knowledge of voiced regions and an estimate of pitch period in advance. The amplitude and frequency modulated (AM-FM) signal model based approach for GCI detection from speech signals was proposed in [22]. The inherent filtering property of the Fourier–Bessel

(FB) series expansion was used to weaken the effects of formants. Peaks of amplitude envelope obtained by applying discrete energy separation algorithm (DESA) to the LPF speech signal were recognized as GCIs. However, this method is applicable only when the filtered speech signal is a mono-component AM-FM signal.

Most of the existing methods require knowledge of voiced parts of speech signal and need a rough estimation of pitch frequency. Pitch frequency is a time varying quantity and does not remain constant throughout complete speech signal. These methods rely on voiced activity detection (VAD) algorithms and pitch frequency estimation algorithms to provide the required information. Moreover, many of the available methods are either sensitive to noise or do not give accurate epoch locations; especially for female speakers, nasals and semivowels. A new method is proposed in this paper based on the nonstationary nature of the speech signal. Speech signals possess time-varying spectrum. The proposed method employs time-frequency analysis in the LFR to detect voiced regions, estimate pitch frequency and locate epochs simultaneously. Time-frequency analysis using TOR is carried out for LFR to suppress the formants. It has been observed that significant energy around the pitch frequency and its harmonics is present only during voiced regions. At epochs, glottal impedance reduces and the magnitude of rate of change of glottal impedance is high, resulting in a high magnitude of rate of change of signal strength. We propose a novel method that locates epochs at peaks of the derivative corresponding to falling edges of peak negative cycles of the LPF speech signal synthesized from TOR coefficients corresponding to LFR.

This paper is organized as follows: An overview of time-order representation (TOR) is presented in Section 2. In Section 3, speech signal behavior in the LFR is discussed. The proposed method is explained in Section 4. Experimental results are presented in Section 5. Section 6 concludes the paper.

2 Time-Order Representation Based on Short-Time Fourier–Bessel Series Expansion

Nonstationary signals are frequently encountered in real environments like speech, electroencephalogram (EEG), earthquake signals etc. Nonstationary signals have time-varying amplitude and spectrum. The Fourier transform assumes the signal to be stationary and thus it is not suitable for efficient analysis of nonstationary signals. Time-frequency analysis is the most efficient way to characterize nonstationary signals. The time-order representation (TOR) based on short-time Fourier–Bessel (FB) series expansion is one such time-frequency technique which is suitable for analysis and synthesis of nonstationary signals. The TOR and FB series expansion employ aperiodic and exponentially decaying Bessel functions as the basis. Recently TOR and FB series expansion have been successfully applied in diversified areas such as postural stability analysis [24], detection of voice onset time [23], separation of speech formants [28], EEG signal segmentation [26], speech enhancement [7] and speaker identification [6]. The FB series expansion has also been used to reduce cross terms in the Wigner–Ville distribution (WVD) [25].

In this paper, we have employed the TOR for speech analysis and its synthesis in the LFR. More specifically; the energy distribution in the time-frequency plane computed from the TOR has been used to detect voiced regions and estimate pitch frequency for each detected voiced region. The speech signal is then reconstructed in the LFR to locate epochs. Hence, we present brief overview of the FB series expansion and the TOR. The zero order FB series expansion of the band-limited signal $x(t)$ spanning over the time period $(0, T)$ has been defined as follows [31]:

$$x(t) = \sum_{l=1}^{Q} C_l J_0\Big(\frac{\lambda_l t}{T}\Big), \tag{1}$$

where λ_l for $l = 1, \ldots, Q$ are the ascending order positive roots of $J_0(\lambda) = 0$, and the zero-order Bessel functions are represented by $J_0(\frac{\lambda_l t}{T})$ in (1). By exploiting the orthogonality of zero-order Bessel functions, the FB coefficients C_l are computed as follows:

$$C_l = \frac{2}{T^2[J_1(\lambda_l)]^2} \int_0^T t x(t) J_0\Big(\frac{\lambda_l t}{T}\Big) dt, \quad l = 1, 2, \ldots, Q, \tag{2}$$

where $J_1(\lambda_l)$ are the first-order Bessel functions. The range Q of the FB series expansion must be equal to the length of the discrete time signal in order to cover the entire bandwidth of the discrete time signal, i.e., the half of the sampling frequency of the discrete time signal. The range and order of the FB series expansion increases with increase in bandwidth and center frequency of the signal, respectively. The order of the FB series expansion and the frequency are related to each other as follows [26]:

$$\lambda_l = 2\pi F T, \tag{3}$$

where F represents the frequency in Hz. The FB coefficients C_l are unique for a given signal. The TOR has been demonstrated [27] to efficiently separate monocomponents of a multicomponent signal. The TOR $X(t_i, l)$ for a given signal $x(t)$ has been defined as follows:

$$X(t_i, l) = \frac{2}{T^2[J_1(\lambda_l)]^2} \int_{\tau=0}^{T} \tau w(\tau - t_i) x(\tau) J_0\Big(\frac{\lambda_l \tau}{T}\Big) d\tau, \tag{4}$$

where $l = 1, \ldots, Q$ and $i = 1, 2, \ldots$. The window function having finite time support is represented by $w(t)$. The window function is real and an even function of time. The TOR is computed at the time t_i where the window is centered. The energy distribution in the time-order plane $E(t_i, l)$ can be computed from TOR coefficients $X(t_i, l)$ as follows [21]:

$$E(t_i,\, l) = \frac{T^2[J_1(\lambda_l)]^2}{2} X^2(t_i, l), \tag{5}$$

and the energy distribution in the time-frequency plane $E(t_i, F)$ can be obtained from the energy in the time-order distribution $E(t_i, l)$ by using the relation in (3). The signal can be reconstructed from TOR coefficients $X(t_i, l)$ as follows:

$$x(t) = \sum_{l=1}^{Q} \sum_{i} X(t_i, l) w(t - t_i) J_0\Big(\frac{\lambda_l t}{T}\Big). \tag{6}$$

The windowing of the signal leads to a trade-off between time and frequency resolution. For acquiring good time resolution one requires a short duration window, whereas good frequency resolution can be achieved by a long time duration window function.

3 Speech Signal Representation in the Low Frequency Range

Pitch frequency varies according to gender, age, emotion, language etc. For adult males, the pitch frequency lies between 60 and 140 Hz, and for adult females the pitch frequency lies between 150 and 400 Hz. For children the pitch frequency can be as high as 500 Hz. The vocal tract system resonates at certain frequencies called as formants. Formants are defined as spectral peaks of the spectrum of the voiced part of speech signal. Information required by humans to distinguish between vowels can be represented quantitatively by the frequency content of vowel sounds. Voiced regions can be detected accurately from the energy distribution in the LFR. The time-varying behavior of speech signals in the LFR can be well understood by computing the energy distribution in the time-frequency plane in the LFR. The contour plot of the energy distribution of speech signal in the time-order plane computed from TOR coefficients corresponding to the frequency range of 0–500 Hz using (5) is depicted in Figure 1 (b). The Gaussian window of size 512 samples (32 ms) has been employed to compute time-order coefficients at each instant of speech signal. The order range corresponding to the frequency range of 0–500 Hz can be computed using the effective length of the windowed signal and the relation in (3). Significant energy is present around the fundamental frequency at 225 Hz and its second harmonic at 450 Hz during voiced regions and

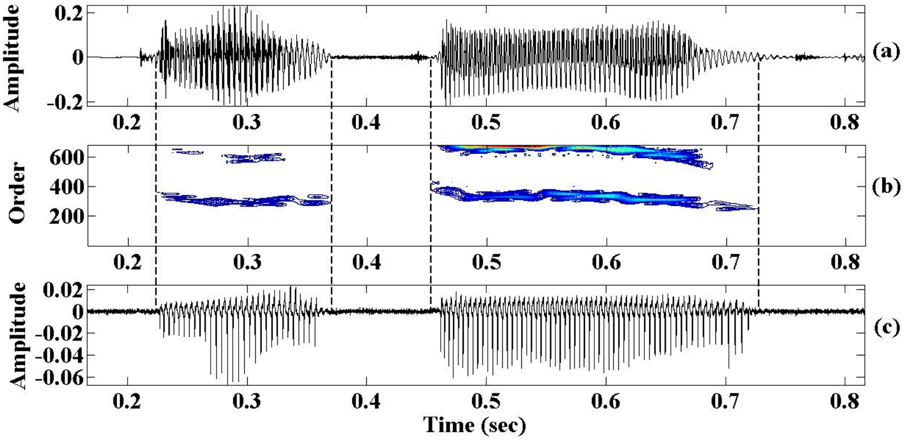

Figure 1. (a) Speech signal segment. (b) Energy distribution of speech signal segment in the time-order plane for LFR. (c) DEGG signal. Reference voiced regions are shown by dashed lines. Speech signal is taken from the CMU-Arctic database with speaker id 30001.

negligible energy is present during non-voiced regions of speech signal. Reference voiced regions are obtained from the differenced EGG (DEGG) signal shown in Figure 1 (c). We propose representation of the discrete time speech signal in the LFR using a multi-component AM-FM signal model as follows:

$$x_{\mathrm{LF}}[n] = \sum_{l=1}^{M} A_l[n] \cos(2\pi l f_0[n]n + \phi_l[n]), \tag{7}$$

where $x_{\mathrm{LF}}[n]$ represents the discrete time speech signal in the LFR. The time varying normalized fundamental frequency (or pitch) is represented by $f_0[n]$. The time varying amplitude envelope and phase of the lth harmonic of the fundamental frequency is represented by $A_l[n]$ and $\varphi_l[n]$, respectively, and M represents the total number of harmonics present in the speech signal in the LFR.

4 Time-Order Representation Based Method for Epoch Detection

Most of the existing methods require knowledge of voiced regions and a rough estimate of pitch frequency for epoch detection from speech signals. Till now, detection of voiced regions, pitch frequency estimation and epoch detection have been treated as independent problems and separate methods have been employed to find solutions to each problem. In this work, we propose a new method based on time-frequency analysis of the speech signal in the LFR to simultaneously detect voiced regions, estimate the pitch frequency for detected voiced regions and accurately identify epochs from the speech signal.

A frequency range of 60–500 Hz is chosen as LFR for time-frequency analysis of the speech signal in order to include the fundamental frequency component, its harmonics and deemphasize formants [11]. Time-order coefficients of the speech signal are computed and the energy density distribution in the time-frequency plane is evaluated using (3) and (5). Marginal energy densities with respect to time and frequency in the LFR are used for detection of voiced regions and an estimation of pitch frequency, respectively. The LPF speech signal is synthesized from the time-order coefficients corresponding to LFR using (6). In order to extract peak negative cycles of LPF speech signal, the fundamental harmonic component of the speech signal is constructed using (6), modified and multiplied with the LPF speech signal. Finally, the derivative of the LPF speech signal corresponding to peak negative cycles of the LPF speech signal is isolated and epochs are identified as peaks of the derivative corresponding to falling edges of peak negative cycles of the LPF speech signal. The proposed method is summarized in the following steps:

1. Time-frequency analysis: Compute discrete time-order coefficients $X_{\mathrm{LF}}[n,l]$ of the discrete time speech signal $x[n]$ for LFR using (4).

2. Energy distribution: Evaluate the energy distribution $E_{\mathrm{LF}}[n,f]$ of the discrete time speech signal $x[n]$ in the time-frequency plane from time-order coefficients $X_{\mathrm{LF}}[n,l]$ corresponding to LFR using (3) and (5).

3. Voiced detection: Perform summation of the energy distribution $E_{\mathrm{LF}}[n,f]$ with respect to frequency to obtain the marginal energy density with respect to time $Z[n]$ in the LFR as follows:

$$Z[n] = \sum_{f \in \mathrm{LFR}} E_{\mathrm{LF}}[n,f]. \tag{8}$$

During voiced activity, substantial energy is present around the fundamental frequency and its harmonics. Vocal regions can be detected from $Z[n]$ by selecting an appropriate threshold. Durations of the speech signal, where the marginal energy density $Z[n]$ is greater than the threshold value, are detected as voiced regions. The speech signal segment and its marginal energy density with respect to time for LFR are shown in Figure 2.

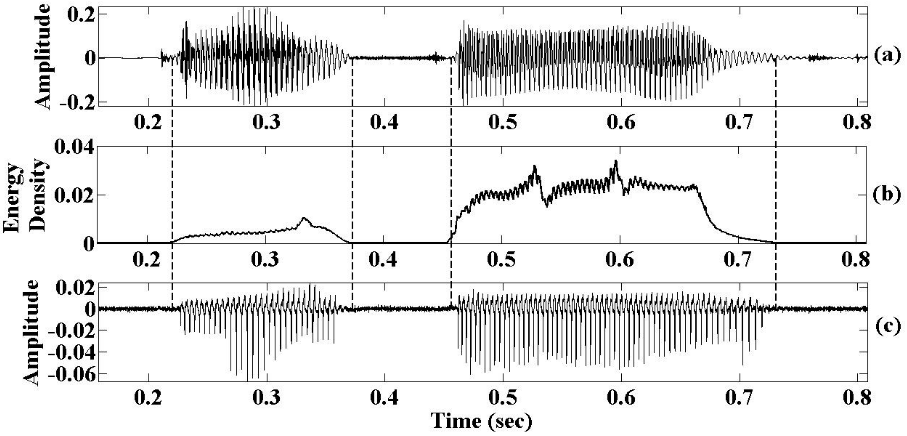

Figure 2. (a) Speech signal segment. (b) Marginal energy density of speech signal segment with respect to time for LFR. (c) DEGG signal. Reference voiced regions are shown by dashed lines. Speech signal is taken from the CMU-Arctic database with speaker id 30001.

4. Pitch frequency estimation: Perform summation of the energy distribution $E_{\mathrm{LF}}[n, f]$ over the duration of each detected voiced region to obtain the marginal energy density with respect to frequency $W_m[f]$ for the mth voiced region as follows:

$$W_m[f] = \sum_{n \in [n_{mL}, n_{mH}]} E_{\mathrm{LF}}[n, f] \quad \text{where } f \in \text{LFR}, \tag{9}$$

where $[n_{mL}, n_{mH}]$ represents the mth voiced region in the speech signal. n_{mL} and n_{mH} denote the detected lower time limit and upper time limit of the mth voiced region, respectively. The marginal energy density function with respect to frequency for the mth voiced region $W_m[f]$ will have peaks occurring at the pitch frequency and its harmonics. The frequency corresponding to the first local maxima of $W_m[f]$ in the LFR provides an estimate of the pitch frequency $F_{\mathrm{PF},m}$ for the mth voiced region as follows:

$$\begin{aligned} W_m[f_1] &= \mathrm{FLM}(W_m[f]), \\ F_{\mathrm{PF},m} &= f_1 \times F_s, \end{aligned} \tag{10}$$

where F_s denotes the sampling frequency and the operator FLM denotes the first local maximum. Frequencies below 60 Hz are not considered as the maximum possible pitch period is 16 ms (see [11]). Marginal energy densities with respect to frequency for two different voiced regions of the speech signal are shown in Figure 3.

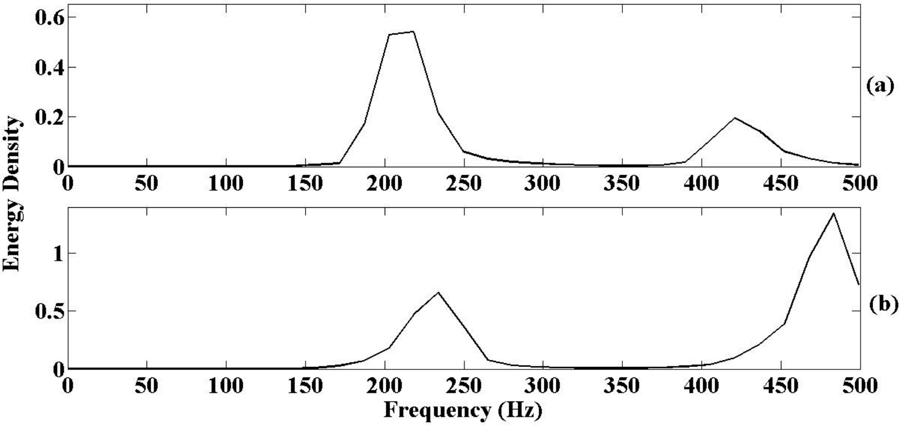

Figure 3. Marginal energy density with respect to frequency for two different voiced regions of the speech signal. Speech signal is taken from the CMU-Arctic database with speaker id 30001.

5. Epoch detection: Epochs are identified as peaks of the derivative corresponding to falling edges of peak negative cycles of the LPF speech signal. The following operations have to be performed to extract epochs automatically:

(a) LPF speech signal: Synthesize the discrete LPF speech signal $x_{\mathrm{LF}}[n]$ from discrete time-order coefficients $X_{\mathrm{LF}}[n,l]$ using (6).

(b) Derivative of the LPF speech signal: Compute the differenced LPF speech signal $x'_{\mathrm{LF}}[n]$ as follows:

$$x'_{\mathrm{LF}}[n] = x_{\mathrm{LF}}[n] - x_{\mathrm{LF}}[n-1] = \Delta x_{\mathrm{LF}}[n], \tag{11}$$

where Δ denotes the difference operator.

(c) Bandwidth of the fundamental harmonic component: For each detected voiced region, a range of frequencies around the fundamental frequency (or pitch) with significant magnitude for the marginal energy density with respect to frequency is taken into account for the reconstruction of the fundamental harmonic component from the time-order coefficients $X_{\mathrm{LF}}(n,l)$ in the LFR. Let $[l_{mL}, l_{mH}]$ denote the range of order of the FB series expansion around the order corresponding to the fundamental frequency $F_{\mathrm{PF},m}$ for the mth voiced region which has significant magnitude for the marginal energy density with respect to frequency $W_m[f]$. l_{mL} and l_{mH} denote the index corresponding to the lower and upper limit of the order range around the fundamental frequency $F_{\mathrm{PF},m}$, respectively.

(d) Modification of time-frequency coefficients: In order to synthesize the fundamental harmonic component from time-order coefficients $X_{\mathrm{LF}}[n, l]$, time-order coefficients $X_{\mathrm{LF}}[n, l]$ corresponding to higher harmonics must be made zero for each voiced region as follows:

$$\tilde{X}_{\mathrm{LF}}[n, l] = \begin{cases} X_{\mathrm{LF}}[n, l] & \text{if } l \in [l_{mL}, l_{mH}] \text{ for all } n, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

(e) Fundamental harmonic synthesis: Reconstruct the fundamental harmonic component $a[n]$ from modified time-order coefficients $\tilde{X}_{\mathrm{LF}}(n, l)$ using (6).

(f) Develop a new signal $b[n]$ from the fundamental harmonic component $a[n]$ as follows:

$$b[n] = \begin{cases} 0 & \text{if } a[n] \geq 0, \\ 1 & \text{if } a[n] < 0. \end{cases} \tag{13}$$

(g) Construct a new signal $c[n]$ from the LPF speech signal $x_{\mathrm{LF}}[n]$ as follows:

$$c[n] = \begin{cases} 0 & \text{if } x_{\mathrm{LF}}[n] \geq 0, \\ x_{\mathrm{LF}}[n] & \text{if } x_{\mathrm{LF}}[n] < 0. \end{cases} \tag{14}$$

(h) Generate a new signal $d[n]$ from the differenced LPF speech signal $x'_{\mathrm{LF}}[n]$ as follows:

$$d[n] = \begin{cases} 0 & \text{if } x'_{\mathrm{LF}}[n] \geq 0, \\ x'_{\mathrm{LF}}[n] & \text{if } x'_{\mathrm{LF}}[n] < 0. \end{cases} \tag{15}$$

(i) Peak negative cycles extraction: Perform multiplication of $b[n]$ and $c[n]$ to isolate peak negative cycles of the LPF speech signal $x_{\mathrm{LF}}[n]$ as follows:

$$g[n] = b[n] \times c[n]. \tag{16}$$

(j) Create a new signal $p[n]$ from $g[n]$ as follows:

$$p[n] = \begin{cases} 0 & \text{if } g[n] \geq 0, \\ 1 & \text{if } g[n] < 0. \end{cases} \tag{17}$$

(k) Generate a new signal $q[n]$ to extract the derivative corresponding to falling edges of peak negative cycles of the LPF speech signal $x_{\mathrm{LF}}[n]$ as follows:

$$q[n] = p[n] \times d[n]. \tag{18}$$

(l) Locate epochs: Time instants corresponding to local minima of $q[n]$ are detected as epochs.

5 Experimental Results

The proposed method has been simulated on speech signals taken from the CMU-Arctic database [5, 13] for performance assessment and comparison of results with the ZFR method. The CMU-Arctic database consists of around 1150 phonetically balanced sentences of about 3 seconds duration carefully selected from out-of-copyright texts from project Gutenberg. They are spoken by five male speakers and two female speakers. It also consists of simultaneous recordings of EGG signals for two male speakers and one female speaker. Speech and EGG signals are digitized in time at a sampling rate of 32 kHz and 16 bit resolution. Time alignment of speech signals and EGG signals was already done in the CMU-Arctic database to compensate for the larynx to microphone delay which was determined to be 0.7 ms.

The time-order representation (TOR) has been used as a time-frequency analysis and synthesis tool. In order to reduce the computational complexity, speech signals obtained from the database have been downsampled to 16 kHz sampling rate. The Gaussian window of size 512 samples (32 ms) has been employed to compute time-order coefficients at each sample instant of the speech signal. The window size cannot be made arbitrarily small for the speech segment spectral properties to have correlation with the spectral properties of the original speech signal. A threshold equal to 1% of the maximum marginal energy density with respect to time is chosen to identify voiced activity regions. Figure 4 shows the speech signal segment of a female speaker, the LPF speech signal, the derivative of the LPF speech signal, the fundamental harmonic component, extracted peak negative cycles, the isolated derivative corresponding to falling edges of peak negative cycles of the LPF speech signal and the DEGG reference signal. Reference epochs are extracted by finding peaks in the DEGG signal during voiced regions. Figure 5 depicts the same signals as mentioned above obtained during the simulation of the proposed method on a speech signal segment of a male speaker. The following measures defined in [20] have been used to evaluate the performance of the proposed method.

1. Larynx cycle: The range of samples $\frac{n_{r-1}+n_r}{2} \leq n \leq \frac{n_r+n_{r+1}}{2}$, given an epoch reference at sample n_r, with preceding and succeeding epoch references at samples n_{r-1} and n_{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.

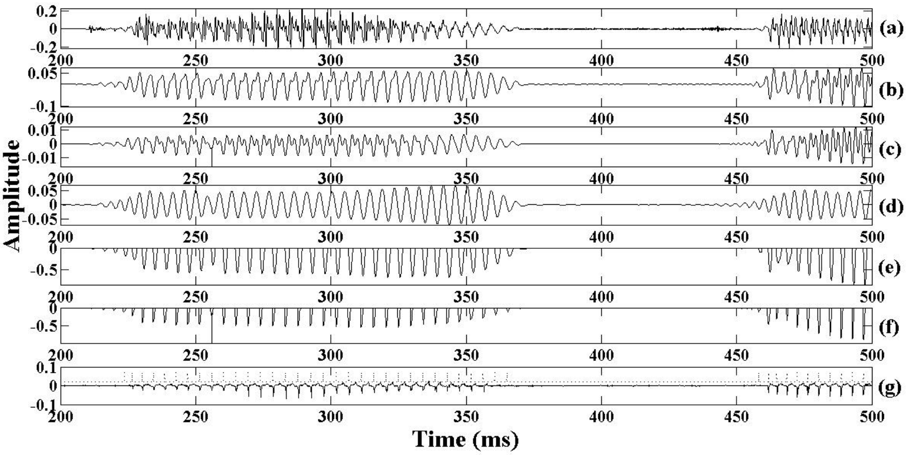

Figure 4. (a) Female speech signal segment. (b) LPF speech signal. (c) Derivative of LPF speech signal. (d) Fundamental harmonic component. (e) Extracted peak negative cycles of LPF speech signal. (f) Derivative corresponding to falling edges of peak negative cycles of LPF speech signal. (g) DEGG reference signal. Pulses in dashed lines correspond to detected epochs. Speech signal is taken from the CMU-Arctic database with speaker id 30001.

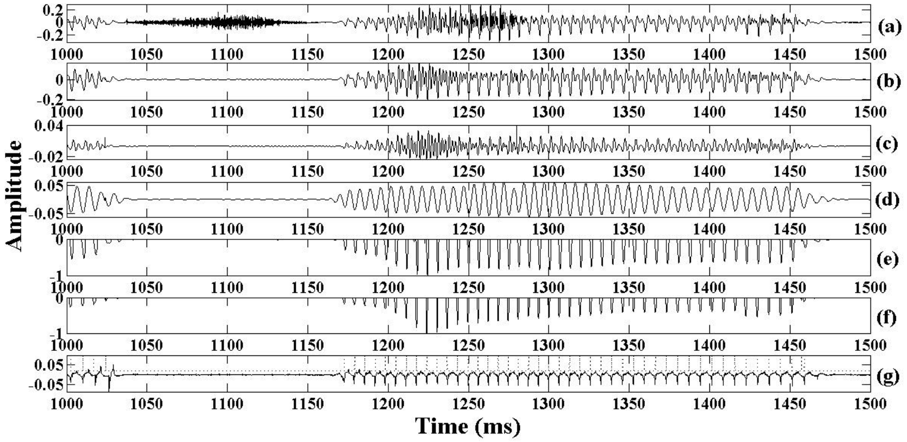

Figure 5. (a) Male speech signal segment. (b) LPF speech signal. (c) Derivative of LPF speech signal. (d) Fundamental harmonic component. (e) Extracted peak negative cycles of LPF speech signal. (f) Derivative corresponding to falling edges of peak negative cycles of LPF speech signal. (g) DEGG reference signal. Pulses in dashed lines correspond to detected epochs. Speech signal is taken from the CMU-Arctic database with speaker id 10201.

Method	IDR (%)	MR (%)	FAR (%)	Mean Absolute IE (ms)	IDA (ms)
Proposed	98.55	0.80	0.65	0.24	0.22
Zero Frequency Resonator	97.96	1.04	0.77	0.46	0.25

Table 1. Performance comparison of epoch detection methods on three male and three female speech signals taken from the CMU-Arctic database.

5. Identification error ζ (IE): The timing error between the reference epoch location and the detected epoch location in larynx cycles for which exactly one epoch was detected. Smaller values of ζ indicate high accuracy of identification.

6. Identification accuracy σ (IDA): It is defined as the standard deviation of the identification error ζ. Smaller values of σ indicate high accuracy of identification.

Table 1 shows the comparison of performance results of the proposed method and the ZFR method on three male and three female speech signals taken from the CMU-Arctic database. The ZFR method [18] requires prior knowledge of voiced regions and an estimate of the average pitch frequency. The proposed method has no such prerequisite and can be employed on entire speech signal. The mean of the absolute value of the identification error for the proposed method is significantly better than the ZFR method. It implies that epochs detected by the proposed method are close to reference epochs identified from the DEGG signal. The performance is nearly the same on all other parameters.

6 Conclusion

In this paper, we have presented a time-frequency analysis based novel method for epoch detection that can be applied on entire speech signal without prerequisite of voiced detection. The proposed method does not depend on the modeling of the vocal tract system. The method has exploited the behavior of the speech signal in the low frequency range (LFR). The marginal energy density with respect to time has been used to identify voiced regions. The fundamental harmonic component extracted from the speech signal has been used to isolate the derivative corresponding to falling edges of peak negative cycles of the LPF speech signal which was synthesized from the time-order coefficients of the speech signal corresponding to the LFR. Epochs were identified as peaks of the isolated derivative signal. The performance of the proposed method has been evaluated on speech signals taken from the CMU-Arctic database. The proposed method has provided excellent results in terms of low mean absolute identification error which enables accurate epoch detection.

Bibliography

[1] T. Ananthapadamanabha and B. Yegnanarayana, Epoch extraction from linear prediction residual for identification of closed glottis interval, *IEEE Trans. Acoust. Speech Signal Process.* **27** (1979), 309–319.

[2] B. S. Atal and S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, *J. Acoust. Soc. Amer.* **50** (1971), 637–655.

[3] M. Brookes, P. A. Naylor and J. Gudnason, A quantitative assessment of group delay methods for identifying glottal closures in voiced speech, *IEEE Trans. Audio Speech Lang. Process.* **14** (2006), 456–466.

[4] Y. M. Cheng and D. O'Shaughnessy, Automatic and reliable estimation of glottal closure instant and period, *IEEE Trans. Acoust. Speech Signal Process.* **37** (1989), 1805–1815.

[5] CMU-Arctic speech synthesis databases, `http://festvox.org/cmu_arctic/index.html`

[6] K. Gopalan, T. R. Anderson and E. J. Cupples, A comparison of speaker identification results using features based on cepstrum and Fourier–Bessel expansion, *IEEE Trans. Speech Audio Process.* **7** (1999), 289–294.

[7] F. S. Gurgen and C. S. Chen, Speech enhancement by Fourier–Bessel coefficients of speech and noise, *IEE Proc. Comm. Speech Vis.* **137** (1990), 290–294.

[8] C. Hamon, E. Mouline and F. Charpentier, A diphone synthesis system based on time-domain prosodic modifications of speech, *Proc. IEEE Int. Conf. Acoustic Speech Signal Process.* **1** (1989), 238–241.

[9] C. Hoelper, A. Frankort, C. Erdmann and P. Vary, A novel voiced/unvoiced/silence classification scheme for offline speech coding, *Proc. European Signal Processing Conf.* **3** (2002), 121–124.

[10] Y. Hu, D. Wu and A. Nucci, Pitch-based gender identification with two stage classification, *Security Comm. Networks* **5** (2012), 211–225.

[11] H. Huang and J. Pan, Speech pitch determination based on Hilbert–Huang transform, *Signal Process.* **86** (2006), 792–803.

[12] P. Jinachitra, Glottal closure and opening detection for flexible parametric voice coding, *Proc. Interspeech* (2006).

[13] J. Kominek and A. Black, The CMU-Arctic speech databases, *5th ISCA Speech Synthesis Workshop* (2004), 223–224.

[14] S. G. Koolagudi, R. Reddy and K. S. Rao, Emotion recognition from speech signal using epoch parameters, *Proc. IEEE Int. Conf. Signal Process Comm.* (2010), 1–5.

[15] A. Kounoudes, P. A. Naylor and M. Brookes, The DYPSA algorithm for estimation of glottal closure instants in voiced speech, *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* **1** (2002), 349–352.

[16] C. Ma, Y. Kamp and L. F. Willems, A Frobenius norm approach to glottal closure detection from the speech signal, *IEEE Trans. Speech Audio Process.* **2** (1994), 258–265.

[17] E. Moore, M. Clements, J. Peifer and L. Weisser, Investigating the role of glottal features in classifying clinical depression, *Proc. IEEE 25th Int. Conf. Eng. Medicine Biology Society* **3** (2003), 2849–2852.

[18] K. S. R. Murty and B. Yegnanarayana, Epoch extraction from speech signals, *IEEE Trans. Audio Speech Lang. Process.* **16** (2008), 1602–1613.

[19] J. L. Navarro-Mesa, E. Lleida-Solano and A. Moreno-Bilbao, A new method for epoch detection based on the Cohen's class of time frequency representations, *IEEE Signal Process. Letters* **8** (2001), 225–227.

[20] 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.* **15** (2007), 34–43.

[21] R. B. Pachori, Discrimination between ictal and seizure-free EEG signals using empirical mode decomposition, *Res. Lett. Signal Process.* (2008), article ID 293056.

[22] R. B. Pachori and S. V. Gangashetty, AM-FM Model based approach for detection of glottal closure instants, *Proc. IEEE 10th Int. Conf. Inf. Sci. Signal Process. Appl.* (2010), 266–269.

[23] R. B. Pachori and S. V. Gangashetty, Detection of voice onset time using FB expansion and AM-FM model, *Proc. IEEE 10th Int. Conf. Inf. Sci. Signal Process. Appl.* (2010), 149–152.

[24] R. B. Pachori and D. Hewson, Assessment of the effects of sensory perturbations using Fourier–Bessel expansion method for postural stability analysis, *J. Intell. Syst.* **20** (2011), 167–186.

[25] R. B. Pachori and P. Sircar, A new technique to reduce cross terms in the Wigner distribution, *Digital Signal Process.* **17** (2007), 466–474.

[26] R. B. Pachori and P. Sircar, EEG signal analysis using FB expansion and second-order linear TVAR process, *Signal Process.* **88** (2008), 415–420.

[27] R. B. Pachori and P. Sircar, Time-frequency analysis using time-order representation and Wigner distribution, *Proc. IEEE Conf. TENCON* (2008), 1–6.

[28] R. B. Pachori and P. Sircar, Analysis of multicomponent AM-FM signals using FB-DESA method, *Digital Signal Process.* **20** (2010), 42–62.

[29] M. D. Plumpe, T. F. Quatieri and D. A. Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification, *IEEE Trans. Speech Audio Process.* **7** (1999), 569–586.

[30] K. S. Rao and B. Yegnanarayana, Prosody modification using instants of significant excitation, *IEEE Trans. Audio Speech Lang. Process.* **14** (2006), 972–980.

[31] J. Schroeder, Signal processing via Fourier–Bessel series expansion, *Digital Signal Process.* **3** (1993), 112–124.

[32] R. Smits and B. Yegnanarayana, Determination of instants of significant excitation in speech using group delay function, *IEEE Trans. Speech Audio Process.* **3** (1995), 325–333.

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

[34] H. W. Strube, Determination of the instant of glottal closure from the speech wave, *J. Acoust. Soc. Amer.* **56** (1974), 1625–1629.

[35] V. N. Tuan and C. d'Alessandro, Robust glottal closure detection using the wavelet transform, *Proc. European Conf. Speech Technology* (1999), 2805–2808.

[36] D. Veeneman and S. BeMent, Automatic glottal inverse filtering from speech and electroglottographic signals, *IEEE Trans. Acoust. Speech Signal Process.* **33** (1985), 369–377.

[37] B. Yegnanarayana and K. S. R. Murty, Event-based instantaneous fundamental frequency estimation from speech signals, *IEEE Trans. Audio Speech Lang. Process.* **17** (2009), 614–624.

[38] B. Yegnanarayana and R. L. H. M. Smits, A robust method for determining instants of major excitations in voiced speech, *Proc. Int. Conf. Acoust. Speech Signal Process.* **1** (1995), 776–779.

[39] B. Yegnanarayana and R. N. J. Veldhuis, Extraction of vocal-tract system characteristics from speech signals, *IEEE Trans. Speech Audio Process.* **6** (1998), 313–327.

Received November 18, 2011.

Author information

Pooja Jain, School of Engineering, Indian Institute of Technology Indore,
Indore-452017, India.
E-mail: poojaj@iiti.ac.in

Ram Bilas Pachori, School of Engineering, Indian Institute of Technology Indore,
Indore-452017, India.
E-mail: pachori@iiti.ac.in