Circuits Syst Signal Process (2015) 34:2309–2333
DOI 10.1007/s00034-014-9957-4

Epoch Extraction Using Zero Band Filtering from Speech Signal

K. T. Deepak · S. R. M. Prasanna

Received: 20 February 2014 / Revised: 8 December 2014 / Accepted: 9 December 2014 /
Published online: 25 December 2014

Abstract Zero frequency filter (ZFF) is a marginally stable infinite impulse response resonant filter at 0 Hz that is used to extract the epoch locations reliably from speech signals. However, the output of such an ideal resonator is an exponentially increasing/decreasing function of time. The trend is removed from the filtered output by subtracting the average over 1–2 pitch periods to obtain zero frequency filtered signal. Alternatively in this paper, a bounded input bounded output stable realization of ZFF is proposed for epoch extraction, where the output of such a filter is not an increasing/decreasing function of time. The advantages of using such a stable filter is that the filter output is bounded and has no precision related problem associated with the output for lengthy speech files, also, the method does not require remove trend procedure that needs initial pitch estimation. The proposed approach is evaluated using CMU-Arctic database for clean and degraded conditions. Furthermore, the method is also validated in cases of singing voice and emotional speech to demonstrate the robustness for varying pitch scenarios. The proposed method is found to be robust for wide range of chosen parameters.

Keywords Epoch extraction · Glottal closure instants (GCI) · Zero frequency filter (ZFF) · Zero band filter (ZBF)

1 Introduction

Epochs are instants of significant excitation present within a pitch period. Most of the time instants of significant excitation takes place during the glottal closure of thevoiced

K. T. Deepak · S. R. M. Prasanna (✉)
Indian Institute of Technology Guwahati, Guwahati, India
e-mail: prasanna@iitg.ernet.in

K. T. Deepak
e-mail: deepakkt@iitg.ernet.in

speech regions [11,18]. Due to time varying nature of both voiced excitations and vocal tract characteristics, estimating accurate location of epochs in voiced regions is still a challenging task. When glottis closes suddenly, a puff of air excites the vocal tract system. During such events, a rapid change takes place in speech signal that gets manifested as sharp peaks in amplitude. However, it is not easy to detect such locations directly from speech signals. It can be imagined that [3] nature of each of the epochs is impulse-like. The train of such impulse-like excitations with varying time and amplitudes gets convolved with time varying vocal tract system. Since, neither source nor vocal tract system is known a priori, separating one of them from other essentially turns out to be a blind deconvolution problem. Knowing accurate location of epochs in speech signal has several applications in speech analysis [24], pitch synchronous-based speech synthesis and desired foreground speech segmentation from rest of the background noise [5]. One of the obvious derivative of knowing accurate epoch locations is the estimation of [23] fundamental frequency (F_0). The F_0 parameter is used in applications like speech synthesis [25] and prosody modification [9]. It is found that the nature of ZFFS offers discriminative information between foreground speech and background noise that includes background speech [5]. It is shown that epoch locations and strength of excitation parameters are useful in segmenting foreground speech regions from rest of the background regions in noisy environments collected naturally. The foreground speech refers to the content spoken by a speaker close to sensor. While rest of the content is termed as background. Many such applications derived as a result of knowing epoch locations and the challenges involved in accurately estimating those locations has motivated researchers toward addressing this problem.

Most of the existing methods in the literature rely on source filter separation model. Inverse filtering operation on speech signal using estimated vocal tract filter is used to derive LP residual signal. Though it is not possible to extract epoch information directly from LP residual due to random polarities [1], many methods exploit it implicitly or explicitly for epoch extraction [1,6,16,17,20,22]. Deriving LP residual from speech signal is based on the assumption that the analysis window of 20–30 ms is stationary. However, this assumption of stationarity does not hold as both voiced excitation and vocal tract system can be varying within analysis window. Also, such methods rely on higher energy of LP residual signal at epoch locations relative to other regions and this may not be true always. Furthermore, based on the global phase characteristics of speech signal obtained from LP residual, many methods were proposed on the basis of group delay function [12,18]. A detailed quantitative review among top four methods based on group delay techniques were present in [2]. However, reported identification rates and accuracies of group delay based methods are relatively poor when compared to current state of the art methods reported in [3].

To alleviate perfect deconvolution problem associated with LP residual-based methods, a method was proposed in [11]. This method does not depends on the critical ability to deconvolve vocal tract system response from voiced excitations. The method was proposed based on the analysis that significant excitations are impulse-like and their strengths are significantly larger than other regions. Hence, those significant excitations are in the form of train of impulses that excites vocal tract system. The effect of such impulses is spread throughout the bandwidth of the speech signal under analysis in frequency domain. It is therefore evident that the impulse-like excitation

information is present at 0 Hz component as well. The vocal tract system has least effect on significant excitations at 0 Hz frequency component. In order to extract this information at 0 Hz, an integrator is designed using a marginally stable 0 Hz resonator and it is called zero frequency filter (ZFF). Effectively a 4th order resonator realized as two cascaded 2nd order resonators are used to have steeper roll off. The nature of the filter output is either exponentially increasing/decreasing function of time. Zero frequency filtered signal (ZFFS) is obtained by removing the average values within each analysis window. The analysis window size depends on 1–2 pitch periods on average for that particular speaker. The epoch locations are obtained from ZFFS, where, the positive zero crossings exactly coincides with epoch locations. Estimation of initial average pitch period of speaker becomes necessary within one to two pitch periods for accurate estimation of epoch locations. Also, since the output of the filter grows/decays as a polynomial function of time, consequently, exceeds the precision range of the processor that results in noisy output for lengthy files. In order to overcome such issues present in ZFF, we propose a method based on stable infinite impulse response (IIR) resonant filter to extract epochs from speech signal. Since, such a filter allows a narrow band of frequencies around 0 Hz to pass through, we prefer to call this filter as zero band filter (ZBF) and the output of the filter as zero band filtered signal (ZBFS).The filter is stable and the output of filter is not exponentially increasing/decreasing function of time. Hence, it is not imperative to remove the trend from output signal in order to obtain accurate epoch locations. The positive zero crossings of the ZBFS indicate the epoch locations.

The rest of the paper is organized as follows: Sect. 2 explains ZFF method in brief and the design analysis of the proposed ZBF method and its impact on epoch location extraction. Also, the section describes the issues present in proposed method and ways to overcome the same. Section 3 describes the details of the evaluation procedure and results obtained in terms of identification rate (IDR) and accuracies. The robustness of the proposed method in terms of varying pitch scenarios is described in Sect. 4. The summary and conclusions of the present work, and the scope for future work are mentioned in Sect. 5.

2 Zero Band Filtering of Speech

The zero frequency filtering method is briefly described here as it is necessary for explaining the proposed method [11]. Let $s[n]$ be the speech signal where the epochs have to be identified. Difference signal $x[n]$ is obtained from $s[n]$ to minimize any low frequency fluctuations present and is given as

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

The difference signal is passed twice through an ideal resonator at zero frequency given by

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

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n], \tag{3}$$

where $a_1 = -2$ and $a_2 = 1$. This is equivalent of successively integrating the input four times, termed more commonly as filtering at zero frequency. The trend in $y_2(n)$ is removed by subtracting the average over 1–2 pitch periods at each sample. This results in signal $y(n)$

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \tag{4}$$

and is called the zero frequency filtered signal (ZFFS). Here $2N+1$ corresponds to the number of samples in 1–2 pitch periods on average for that particular speaker. Using ZFFS the epoch locations can be exactly located at positive zero crossings.

In order to explain the output to input relationship using a 2nd order filter from Eq. (2) and to represent the system transfer function in generalized form, Eq. (2) can be expanded as

$$y_1[n] = -a_1 y_1[n-1] - a_2 y_1[n-2] + x[n] \tag{5}$$

The z-domain equivalent of Eq. (5) is

$$Y_1(z) = -a_1 Y_1(z) z^{-1} - a_2 Y_1(z) z^{-2} + X(z) \tag{6}$$

and rearranging Eq. (6) we get

$$\frac{Y_1(z)}{X(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{7}$$

generalizing Eq. (7) by substituting $a_1 = -2r$ and $a_2 = r^2$ [15] we get

$$\frac{Y_1(z)}{X(z)} = \frac{1}{1 - 2rz^{-1} + r^2 z^{-2}} \tag{8}$$

where r represents the value of the radius on unit circle in z-plane at which the poles are placed and it's equivalent time domain Equation is given as

$$y_1[n] = x[n] + 2r y_1[n-1] - r^2 y_1[n-2] \tag{9}$$

Equation (9) can be re arranged as follows:

$$y_1[n] - r y_1[n-1] = x[n] + r y_1[n-1] - r^2 y_1[n-2] \tag{10}$$

Let

$$y_3[n] = y_1[n] - ry_1[n-1] \tag{11}$$

substituting Eq. (11) in (10) and rearranging the equation we get

$$y_3[n] = x[n] + r(y_1[n-1] - ry_1[n-2]) \tag{12}$$

using Eq. (11), the above Eq. (12) can be re-written as

$$y_3[n] = x[n] + ry_3[n-1] \tag{13}$$

Equation (13) can be expanded as the following infinite series in terms of input signal with the assumption that $y[n] = 0$ for $n < 0$

$$y_3[n] = x[n] + rx[n-1] + r^2x[n-2] + \ldots \infty \tag{14}$$

Equation (14) can be written as summation series shown as

$$y_3[n] = \sum_{m=0}^{\infty} r^m x[n-m] \tag{15}$$

Equation (11) can be re-written as

$$y_1[n] = y_3[n] + ry_1[n-1] \tag{16}$$

the above Eq. (16) can be written in summation series similar to Eq. (15)

$$y_1[n] = \sum_{l=0}^{\infty} r^l y_3[n-l] \tag{17}$$

substituting Eq. (15) in (17) we get

$$y_1[n] = \sum_{l=0}^{\infty} r^l \sum_{m=0}^{\infty} r^m x[n-m-l] \tag{18}$$

where Eq. (18) can be re-written as

$$y_1[n] = \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} r^{(m+l)} x[n-m-l] \tag{19}$$

If $m + l = k$ in the above Eq. (19), the equation can be written as

$$y_1[n] = \sum_{k=0}^{\infty} (k+1) r^k x[n-k] \tag{20}$$

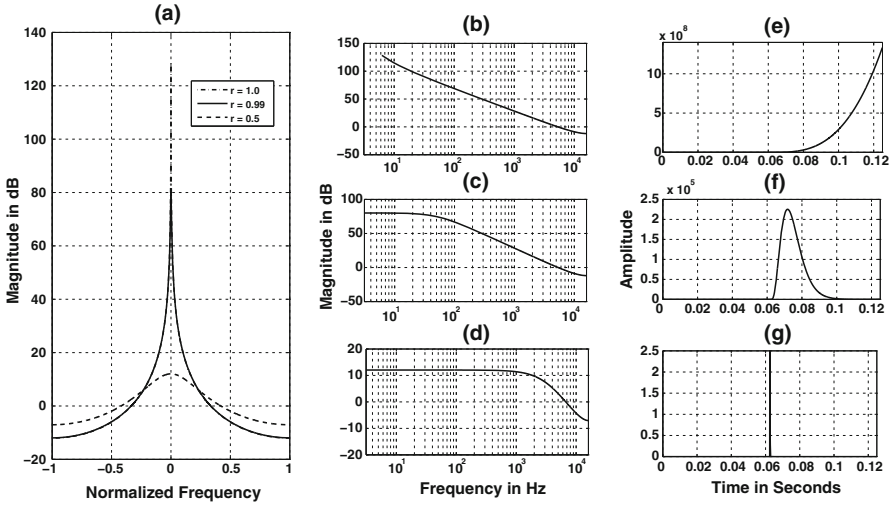

Fig. 1 Second-order filter responses. **a** Magnitude response (normalized frequency axis) when poles are placed at $r = 1.0$, 0.99 and 0.5. Magnitude responses (semilogx) and corresponding impulse responses when poles are placed at $r = 1$ (**b**) and (**e**), $r = 0.99$ (**c**) and (**f**), $r = 0.5$ (**d**) and (**g**)

Equation (20) is the generalized form of representation that is derived in [19]. When $r = 1$, in Eq. (20) represents the ZFF and it is evident that the impulse response of the system is diverging type. This explains the reason for the nature of *zero frequency filter* output growing/decaying exponentially.

2.1 Zero Frequency to Zero Band Filtering

If $r < 1$ the impulse response of the system is converging type. From Eq. (8) varying value of r results in a set of filters having different magnitude responses. This is illustrated by magnitude response plots of 2nd order filter for three different values of r at 1.0, 0.99, and 0.5 in Fig. 1(a). The magnitude response of the filter at $r = 0.99$ is almost same as that of $r = 1.0$ except that it represents bounded case and gain at 0 Hz for $r = 0.99$ is relatively low compared to $r = 1.0$. Furthermore, Fig. 1(b)–(d) represents the magnitude responses of 2nd order filter when poles are placed at $r = 1$, $r = 0.99$, and $r = 0.5$. While Fig. 1 (e)–(g) are corresponding impulse response plots, for clarity purpose, the frequency axis is represented in semi-log scale and an impulse at time instant of 0.0625 s. It can be noticed that as r value increases, the nature of the low pass magnitude response appears to be sharper. Also, the dynamic range of the filter increases as r value increases, thereby increasing the gain value at 0 Hz relative to other frequency components. When, r approaches unity the filter gets manifested as ideal resonator and from Eq. (8) it is evident that the gain of filter is ∞ at 0 Hz. This is achieved at cost of placing the poles on unit circle that results in marginally stable filter. If speech signal is passed through such a filter, the output results in exponentially growing/decaying function of time. While the epochal information is present over the exponential trend that may not be visible directly from the plots. However,

when such an exponential trend is removed from the output signal by using the average over 1–2 pitch periods as shown in Eq. (4), ZFFS signal is obtained. Alternatively, to avoid such an issue we prefer to apply the stable filter by placing poles within the unit circle. Advantage of using such a stable filter is that the output is converging type. The stability is obtained at the cost of finite gain at 0 Hz and an increase in bandwidth of the filter. However, if the filter magnitude response is sufficiently narrow enough that allows a band of frequencies near 0 Hz without affecting the ability to extract epochs then such a filter is preferable. Since, such a stable filter allows a narrow band of frequencies near 0 Hz, we would like to call such a filter as ZBF.

Let h_{ZF2} and h_{ZB2} be the impulse response of 2nd order ZFF and ZBF, respectively. The impulse response h_{ZF2} and h_{ZB2} can be obtained from Eq. (20) as follows:

$$h_{ZF2}[n] = \sum_{k=0}^{\infty}(k+1)\delta[n-k] \tag{21}$$

Equation (21) represents the impulse response of ZFF when $r = 1$.

$$h_{ZB2}[n] = \sum_{k=0}^{\infty}(k+1)r^k\delta[n-k] \tag{22}$$

Equation (22) represents the impulse response of ZBF and $r < 1$ in case of ZBF.

The voiced speech signal can be considered as the convolution between train of impulse excitations and vocal tract system response and it is of interest to observe the filter output response to such train of impulses. Both period and amplitude of such train of impulses vary practically in speech signal. In order to simplify the analysis, the period of such train of impulses is considered to be constant N having unit amplitude. The input signal $x[n]$ in the form of train of impulses is given by Eq. (23) with period N and having unit amplitude.

$$x[n] = \sum_{j=0}^{\infty}\delta[n-Nj] \tag{23}$$

Using Eq. (21) the ZFF output response for train of impulses given in Eq. (23) is given by

$$y_{ZF2}[n] = \sum_{j=0}^{\infty}h_{ZF2}[n-Nj] \tag{24}$$

Similarly, using Eq. (22) the ZBF output is given as

$$y_{ZB2}[n] = \sum_{j=0}^{\infty}h_{ZB2}[n-Nj] \tag{25}$$

When two such filters are cascaded then the impulse response of 4th order ZFF is given as

$$h_{ZF4}[n] = h_{ZF2}[n] * h_{ZF2}[n] \tag{26}$$

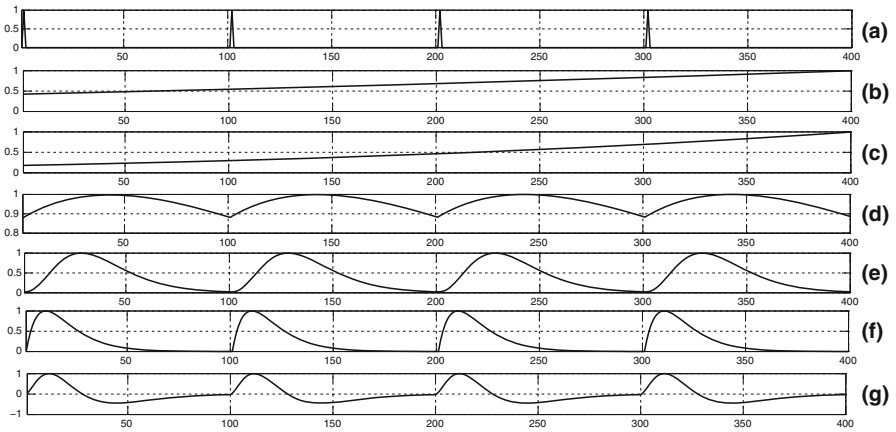

Fig. 2 Illustration of 2nd and 4th order filter output response for input train of impulses when poles are placed at different r values. **a** Input train of impulses and the filter response when poles are placed at (**b**) 2nd order filter output when $r = 1.0$, **b** 4th order filter output when $r = 1.0$, **c** 2nd order filter output when $r = 0.99$, **d** 4th order filter output when $r = 0.99$, **e** 2nd order filter output when $r = 0.9$, **f** 4th order filter output when $r = 0.9$

similarly 4th order impulse response of ZBF is

$$h_{ZB4}[n] = h_{ZB2}[n] * h_{ZB2}[n] \tag{27}$$

The output response of the 2nd order filters are given by Eqs. (24) and (25) to such train of impulses as input when poles are placed at $r = 1$ in case of ZFF, while the poles are placed within unit circle in case of ZBF. The output response can be observed in Fig. 2 for the train of impulses as an input where the period N is 100 samples. It can be observed from Fig. 2(b) and (c) the output is an increasing function of time when poles are placed at $r = 1$ for 2nd and 4th order filters, respectively. However, it can be observed from Fig. 2(d)–(g) that the output is bounded and forms the converging series. It can be further noticed that the output has faster decay when poles are placed farther away from unit circle from Fig. 2(f) and (g), when poles are placed at $r = 0.9$ for 2nd and 4th order filters, respectively.

As, we can notice from Fig. 1 that placement of poles plays a critical role for fixing the magnitude response of the filter. Also, using higher order filters one can increase the gain of the filter at 0 Hz component relative to other components. Hence, we prefer to use a 4th order filter to obtain a better roll off as used in ZFF. Figure 3 illustrates the effect of 4th order filter (realized as cascade of two 2nd order filters) for different pole placement values on the output signal when applied on speech signal from a male speaker in CMU-Arctic database, where a segment of voiced speech region is displayed.

Figure 3(b)–(e) are the 4th order normalized filter outputs for the segment of speech signal for different pole placements at radius r of 0.8, 0.9, 0.99 and 1.0. We can observe from Fig. 3 that as poles are farther away from unit circle the filter allows more frequency components to pass relative to 0 Hz component. Hence, it is preferable to

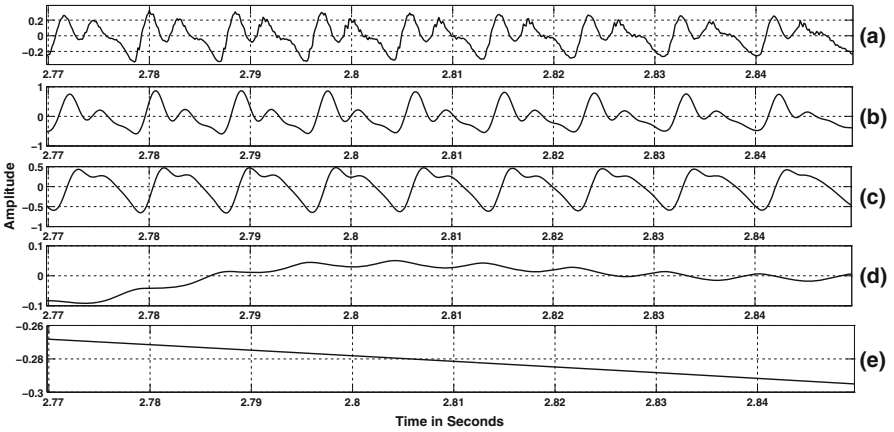

Fig. 3 Illustration of 4th order filter outputs for different values of pole placements. **a** Voiced speech segment of vowel /i/ from a speech file taken from CMU-Arctic database, fourth order filter output of a voiced speech segment in (**a**) when poles are placed at **b** $r = 0.8$, **c** $r = 0.9$, **d** $r = 0.99$, **e** $r = 1.0$

design a filter by placing the poles close to unit circle to avoid interference of vocal tract system response with epochal information. However, if the poles are placed on the unit circle the filter becomes marginally stable, where the output is either exponentially growing or decaying function of time as we can notice this from Fig. 3(e). Hence, to avoid any such issues it is preferable to use a stable filter, while poles are placed close to unit circle.

It can be observed from Fig. 3(d) that when poles are placed at 0.99 the epochal information is overriding on a low frequency fluctuation. This low frequency fluctuations can be eliminated by highpass butterworth filter. In order to have a sharper transition a 4th order filter is preferable. However, the parameters for ZBF and the highpass filter are chosen based on the experiments conducted for varying pole placements and highpass filter cutoff frequencies, respectively. In order to eliminate any bias in the chosen parameters, two different sets of databases belonging to a male and female speakers are selected from CMU-Arctic databases for experimentation. First the experiment involves selection of pole placement by varying its value for ZBF by assessing the performance of the method for fixed highpass cutoff frequency. The performance is measured in terms of IDR, miss rate (MR) and false alarm rate (FAR) for varying values of poles from 0.5 to 0.99 while keeping the highpass filter cutoff frequency at 80 Hz. Figure 4(a), (c) and (e) shows the plots of IDR, MR, and FAR evaluated for two speakers databases. It can be observed that as the pole is farther away from unit circle in the range of 0.5–0.9 IDR of epoch extraction is poor because ZBF allows larger bandwidth relative to 0 Hz component. As a result, the ZBFS has many spurious positive zero crossings that leads to higher false alarms and this is evident from Fig. 4(e). In contrast the IDR improves significantly when the poles are placed in the range of 0.95–0.99. The poles placed at 0.99 is selected because ZBF has the least bandwidth relative to other positions. The highpass filter cutoff frequency is chosen based on experiment conducted by fixing the ZBF pole placed at 0.99 and varying cutoff frequencies from 10 to 100 Hz in steps of 10 Hz. Figure 4(b), (d) and

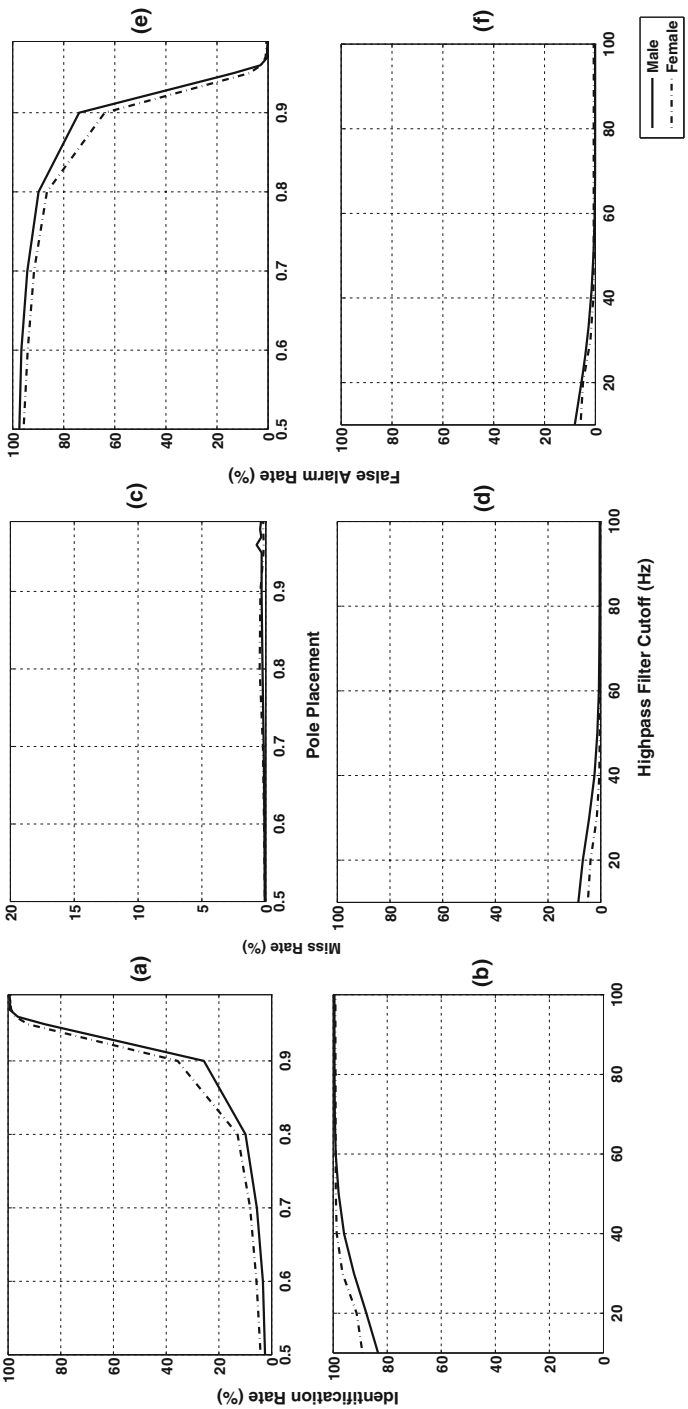

Fig. 4 Performance evaluation of ZBF using variable parameters, **a** and **b** are identification rate, **c** and **d** are miss rate, **e** and **f** are false alarm rate for varying values of pole placement and high pass filter cutoff frequency, respectively

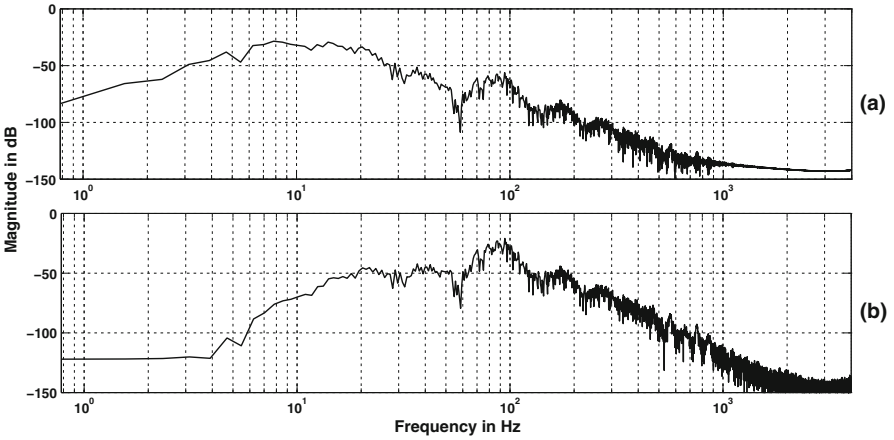

Fig. 5 Spectrum plot of zero band filter output for a speech file from CMU-Arctic database, **a** before and **b** after passing through 4th order butterworth high-pass filter at cutoff frequency of 80 Hz

(f) shows the plots of IDR, MR and FAR evaluated for varying highpass filter cutoff frequencies. It can be observed that IDR is lower when the cutoff frequencies are in the range of 10–60 Hz and improves for the range of 60–100 Hz. Highpass filter cutoff frequency of 80 Hz is selected in the proposed method so that it matches closely with lower pitch range for a male speaker. The Fig. 5(a) and (b) shows the spectrum plots of the filter output for a speech signal before and after applying the high pass filter, respectively. We can notice that the dominant low frequency content is suppressed by a high pass filter at 80 Hz to give emphasis to a band of frequencies that contain epochal information. Figure 6 shows the plots of ZFFS, ZBFS, and the differenced EGG for a voiced speech segment. Figure 6(b) and (c) shows the plots of ZFFS and ZBFS, respectively, for the voiced speech segment shown in Fig. 6(a) and as reference the differential EGG is plotted in Fig. 6(d). As, it can be observed that ZFFS and ZBFS are almost similar and their corresponding positive zero crossings match with reference epoch locations.

Also, from Eq. (20) it is evident that the output of ZFF grows/decays exponentially as function of time. This poses a problem for removing trend from the output to obtain ZFFS in lengthy speech files. The length of files recorded can be of several minutes. Whereas ZBF being a stable filter has no such issues when length of speech files runs for several minutes. This is demonstrated in Fig. 7, where (a) shows the speech waveform of 5 min in length, while (b) and (c) shows ZFFS and ZBFS, respectively, for the same speech signal. The speech file is taken from NIST 2012 mic database [13]. It is observed that the ZFFS is noisy above 1.6 min. This is further demonstrated in Fig. 8 where a segment of voiced speech region is chosen from the same speech waveform that is shown in Fig. 7(a), while Fig. 8(b) and (c) shows ZFFS and ZBFS of the selected segment of voiced speech region, respectively. The voiced segment is selected from the time index of 3.764–3.769 min of the speech file. It can be observed that ZFFS output appears noisy and this may lead to spurious detection of epoch locations because of multiple positive zero crossings. While the ZBFS output has no such issues.

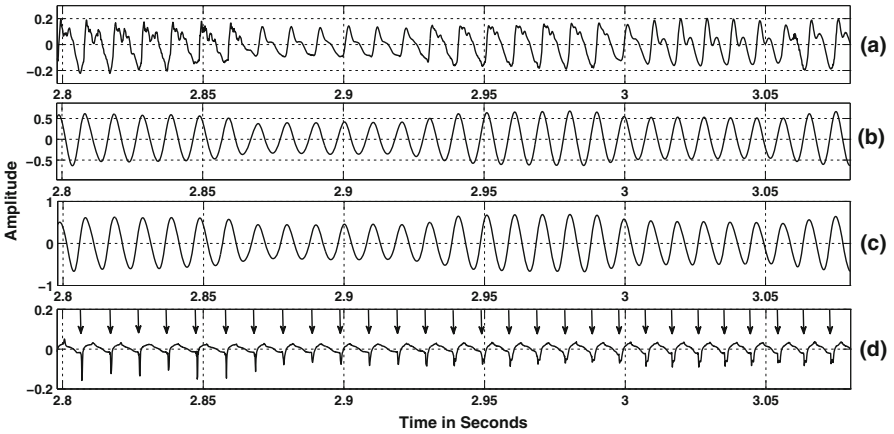

Fig. 6 Comparison of ZFFS and ZBFS. **a** Voiced speech segment of vowel /e/ from a speech file taken from CMU-Arctic database, **b** zero frequency filtered output for voiced speech segment in (**a**), **c** zero band filtered output for voiced speech segment in (**a**), **d** differenced EGG of voiced speech segment shown in (**a**) and *arrows* representing the actual epoch locations

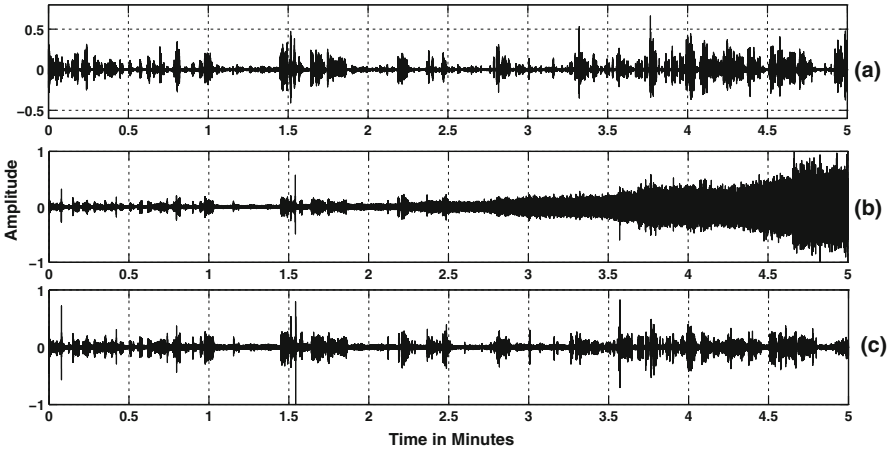

Fig. 7 Illustration of robustness of ZBF for lengthy speech file. **a** Speech waveform from a file taken from mic set of NIST 2012 database, **b** zero frequency filtered output for the waveform shown in (**a**), **c** zero band filtered output for the waveform shown in (**a**)

3 Experiments and Results

The evaluation procedure is carried using clean CMU-Arctic databases and subsequently, the performances are evaluated by degrading the clean speech by different kinds of additive noises at different levels taken from Noisex database. The CMU-Arctic databases are freely available from Festvox website [4]. Three different databases available from the website consists of 2 male and 1 female speakers. All these databases have simultaneous recordings of electroglottograph (EGG) along with speech in two different channels. In order to identify the glottal closure instants from

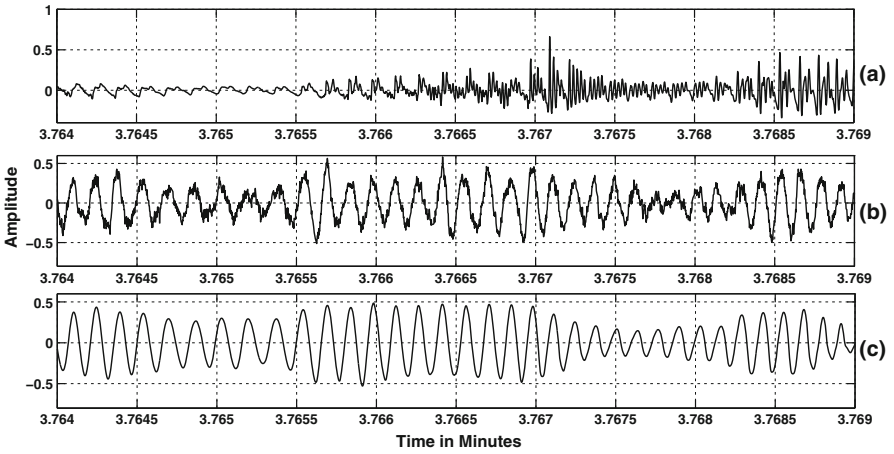

Fig. 8 Illustration of robustness of ZBF by selecting a segment of speech from a lengthy file. **a** Voiced speech segment expanded from Fig. 7(a), **b** zero frequency filtered output and **c** zero band filtered output for the voiced speech segment shown in (false alarm rate)

EGG signal, SIGMA algorithm [21] is used that is available in voicebox and such instants are used as reference markings against which the performance of different methods are compared. Three databases namely BDL-US male, JMK-Canadian male, and SLT-US female have a total of 1,132 phonetically balanced sentences recorded in a closed booth in a controlled environment at the sampling rate of 32 kHz. The performance was evaluated considering the voiced regions from EGG.

The proposed method is evaluated and compared with DYPSA [12], ILPR [16], SEDREAMS [6], YAGA [22] and ZFF [11] in terms of IDR, MR, FAR and identification accuracy (IDA). These parameters are computed as shown in Fig. 9 [12]. The parameters are computed with reference to glottal cycle derived from EGG as reference and is given by $1/2(c_{i-1} + c_i) \leq n \leq 1/2(c_i + c_{i+1})$, and the cycle is defined with respect to i^{th} epoch as reference, denoted by c_i. Identification rate (IDR) represents the percentage of times a single epoch located from method under evaluation within a cycle, MR is when no epochs are located within a cycle and FAR is the number of times multiple epochs detected within a cycle in the database. However, D represents the error from the actual location of the epoch when exactly one epoch is detected by the method under evaluation and IDA is the standard deviation of the error D.

The lengthy speech waveforms are simulated by concatenating many speech files from CMU-Arctic database to evaluate the performance of ZFF and ZBF on lengthy speech files.

3.1 Performance Evaluation on Clean and Degraded Speech

Table 1 shows the results evaluated for clean speech data taken from CMU-Arctic databases. The performance is evaluated in terms of IDR, MR, FAR, and IDA. The performance of the proposed method is compared with five other state of the art

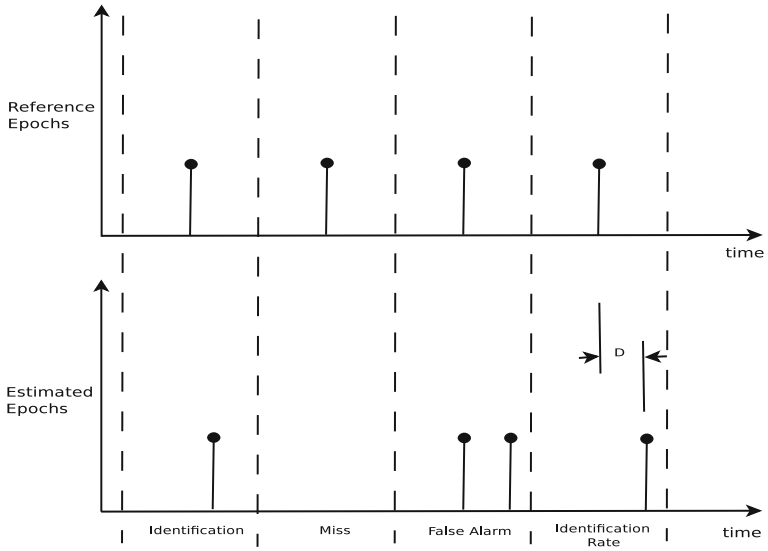

Fig. 9 Characterization of epoch location estimates showing four larynx cycles with examples of each possible outcome from epoch estimation [12]. Identification accuracy is measured by D

Table 1 Performance comparison of epoch extraction methods under clean condition on CMU-Arctic database

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)	IDR (±0.25 ms)
DYPSA	96.66	1.76	1.58	0.59	52.46
ILPR	98.64	0.62	0.71	0.29	86.91
SEDREAMS	97.87	1.14	1.07	0.39	82.59
YAGA	98.71	0.48	0.79	0.31	90.16
ZBF	98.20	0.72	1.06	0.39	86.49
ZFF	99.04	0.18	0.77	0.36	91.26

IDR identification rate, *MR* miss rate, *FAR* false alarm rate, *IDA* identification accuracy

methods. The DYPSA algorithm relies on group delay function which is the average slope of the phase spectrum derived from short time Fourier transform (STFT) of LP residual. The negative zero crossings of group delay function corresponds to epoch locations. However, it is not guaranteed that the epochs always leads to negative zero crossings in the group delay function. The phase slope projection technique is adopted in order to project the slope in between maxima and minima in order to identify the epoch locations. Furthermore the identified epoch locations are pruned based on the distance criteria of adjacent epoch locations detected using dynamic programming [12]. As an further improvement over DYPSA an efficient method to detect epoch locations was suggested using group delay function derived from the LP residual signal. The LP residual signal is derived using non pre-emphasized speech signal rather than commonly used pre-emphasized speech. However, the method makes

use of multi scale product signal derived from wavelets to obtain a train of impulse-like signal. The impulse locations are identified by the zero crossings of group delay function obtained from LP residual signal that is further pruned by N-best dynamic programming [22]. Recently a method was proposed based on the mean-based signal that is derived directly from speech signal. The mean-based signal is derived based on 1.5–2 times the pitch period of the speech signal as an first step which is used to locate the epochs. However, the epoch locations are further refined using residual excitation for accurate estimation in second step in [6]. More recently a method was proposed based on inverse filtering of non pre-emphasized speech called as integrated linear prediction residual (ILPR) using which the half-wave rectified ILPR is derived. Such a signal closely corresponds to train of impulses and those impulses closely match with epoch locations. In order to identify such locations, a temporal measure is adopted to detect the locations of transients called dynamic plosion index [16].

The results shown in Table 1 are the average scores obtained by evaluating the algorithms using three different databases. It can be noticed that the overall IDR and IDR within ± 0.25 ms is better in case of ZFF and YAGA methods compared to other methods. However, the performance of the proposed method is comparable with ILPR in terms IDR and IDR within ± 0.25 ms. The reason for lower IDR and IDA of the proposed method compared to ZFF may be due to relatively larger passband of ZBF at 0 Hz. In terms of IDA, ILPR-based method is better with an average deviation of 0.29 ms compared to other methods. While both SEDREAMS and ZBF are comparable in terms of IDA. Furthermore to evaluate the robustness of the proposed method the performance is computed by adding different kinds of noises to clean speech files taken from CMU-Arctic databases. The results are evaluated for two different noisy conditions, namely white and babble noise taken from Noisex-92 database [14]. Noise is added to clean speech at six different levels of 40, 30, 20, 10, 5, and 0 dB.

Figure 10(a)–(d) shows the performance evaluation of DYPSA, ILPR, SEDREAMS, YAGA, ZFF, and ZBF for degraded conditions by adding white noise at different levels to clean speech files in terms of IDR, MR, FAR, and IDA. The scores plotted in Fig. 10 are the average scores obtained from all speakers considered. It can be noticed that the performance of ZFF, ZBF , SEDREAMS and ILPR-based methods are highly robust to additive white noise in terms of IDR. However, there is considerable decrease in the performance of YAGA in terms of IDR for higher levels of noise added. Furthermore, the IDA of ZFF, ZBF, and ILPR remains robust for higher levels of white noise added. It can be observed that the performance of DYPSA is the most affected by the degradation of speech files by additive white noise in terms of IDR and IDA. One of the important types of noise that can interfere with the speech signal is babble noise and it is a challenge for any speech processing methods as the noise can be spectrally similar to speech signal itself. Hence, the performance of different methods is evaluated for additive babble noise at various levels. Figure 11(a)–(d) shows the performance evaluation of different methods at 40, 30, 20, 10, 5, and 0 dB levels. It can be observed that the performance of ZFF, ZBF, ILPR, and YAGA is relatively robust at higher levels of additive babble noise compared to other methods. Though there is a slight degradation of performance of ZBF relative to ZFF at higher levels of additive babble noise, the performance of the proposed method remains robust for high levels of additive babble noise relative to other methods. The

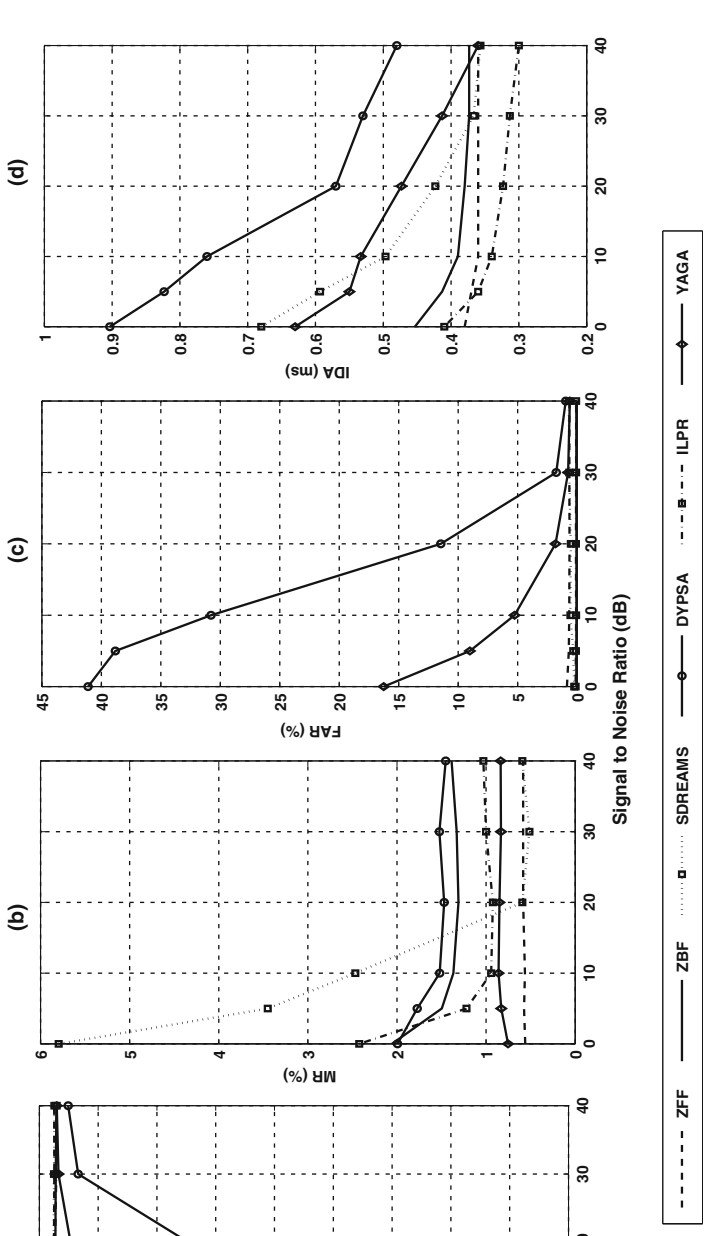

Fig. 10 Performance comparison of epoch extraction under degraded conditions by adding white from Noisex-92 database to CMU-Arctic database (average scores of three different speakers) at 40, 30, 20, 10, 5 and 0 dB levels. **a** IDR, **b** MR, **c** FAR, **d** IDA for additive white noises

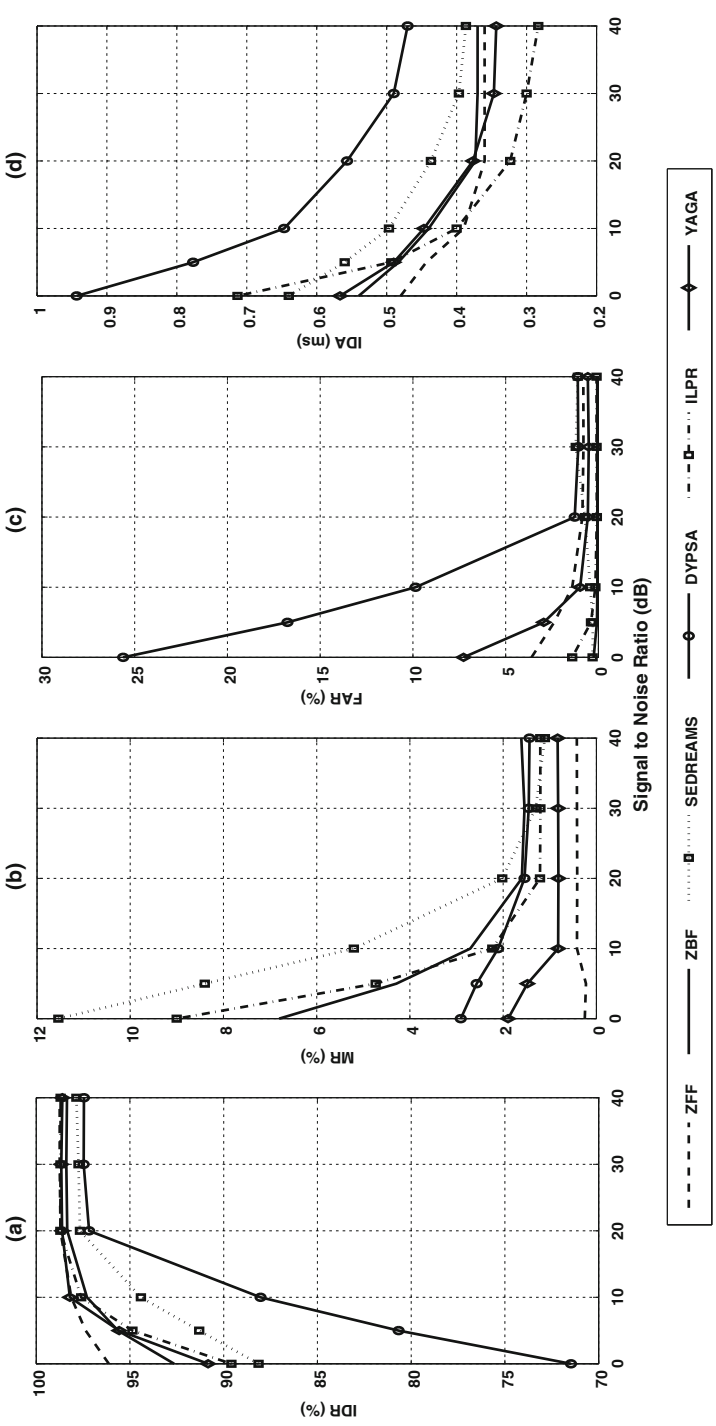

Fig. 11 Performance comparison of epoch extraction under degraded conditions by adding babble from Noisex-92 database to CMU-Arctic database (average scores of three different speakers) at 40, 30, 20, 10, 5 and 0 dB levels. **a** IDR, **b** MR, **c** FAR, **d** IDA for additive babble noises

overall performance of ZBF remains robust for additive white and babble noises at different levels.

3.2 Performance Evaluation on Lengthy Speech Waveform

In order to demonstrate the reliability of the proposed method for lengthy speech cases, two different lengthy speech data of approximately 10 min were considered for each of male and a female speaker from CMU-Arctic databases. The lengthy speech data were simulated by concatenating 200 speech files added with 20 dB white noise from each of the speaker's database, respectively. Since, each of the speech files in CMU-Arctic database is approximately 3 s in length, 200 files are concatenated to form an approximately 10 min speech data. Such a lengthy speech data is passed through ZFF and ZBF to obtain ZFFS and ZBFS, respectively. With prior knowledge of time stamps for each of the individual speech files the performance were evaluated in terms of IDR, MR, FAR and IDA using EGG data as reference. The results obtained are shown in the form of graphs in Fig. 12(a)–(h). Figure 12(a)–(d) shows the plots of IDR, FAR, MR, and IDA, respectively, for male speaker's data and similarly Fig. 12(e)–(h) are the plots of IDR, FAR, MR, and IDA, respectively, for female speaker's data. The scores are normalized and smoothed by moving average filter for clarity. As, it can be noticed that the IDR exponentially falls in case of ZFF with increase in the length of speech, however, the performance of proposed method is robust for arbitrary lengths of data. The reason for decreased IDR in case of ZFF can be accounted by exponential increase in the FAR due to spurious detection of positive zero crossings in ZFFS as mentioned in Sect. 2. However, it can be observed that MR does not increase with the length of speech waveform in case of ZFF. Also, IDA decreases with the length of the speech waveform in case of ZFF while ZBF remains robust. It can be observed that the performance of the ZFF decreases approximately starting from 0.8 min in case of male speaker, while it starts decreasing from 3 min in case of female speaker. This may be explained from Eq. (20) for $r = 1$ that the integrator output depends on the nature of input signal along with the increase in time index.

4 Robustness of Epoch Extraction using ZBF

The robustness of ZBF is further demonstrated in cases of foreground speech [5], singing voices and emotional speech. Although, ZBF allows a narrow band of frequencies around 0 Hz relative to ZFF, it is shown that ZBF is equally robust to additive noises in Sect. 3.1. However, the nature of degradation of speech collected from natural environments is different from additive noises. Hence, the performance of different methods is compared for speech files collected from different natural environments. Also, ZBF does not require a priori pitch estimation and it is of interest to evaluate the performance for varying pitch scenarios as in cases of singing voices and emotional speech.

4.1 Epoch Extraction for Foreground Segmentation

In present day scenario there is no restriction on the environment from which the users can access speech recognition and speaker verification systems. The environments

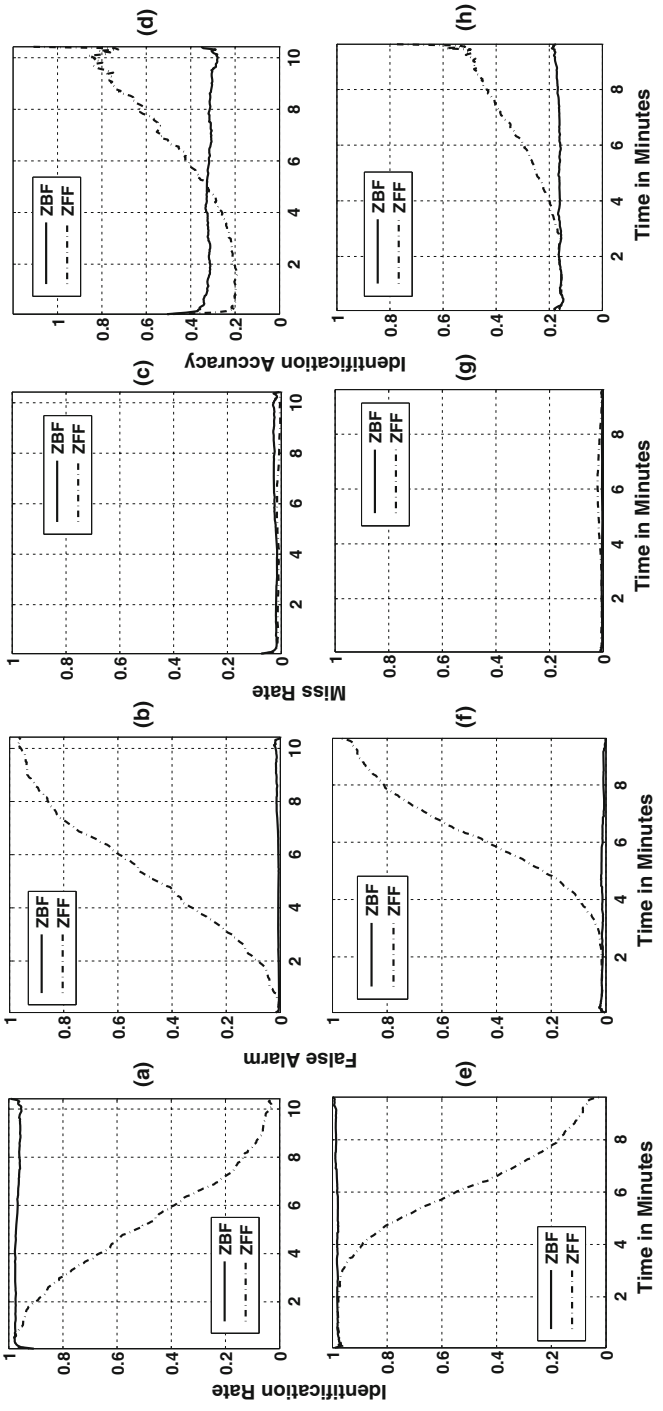

Fig. 12 Performance evaluation of ZFF and ZBF on concatenated lengthy speech waveform generated from 200 files from CMU-Arctic database. **a** Identification rate, **b** false alarm rate, **c** miss rate and **d** identification accuracy, for a male speaker, **e** identification rate, **f** false alarm rate, **g** miss rate and **h** identification accuracy, for a female speaker

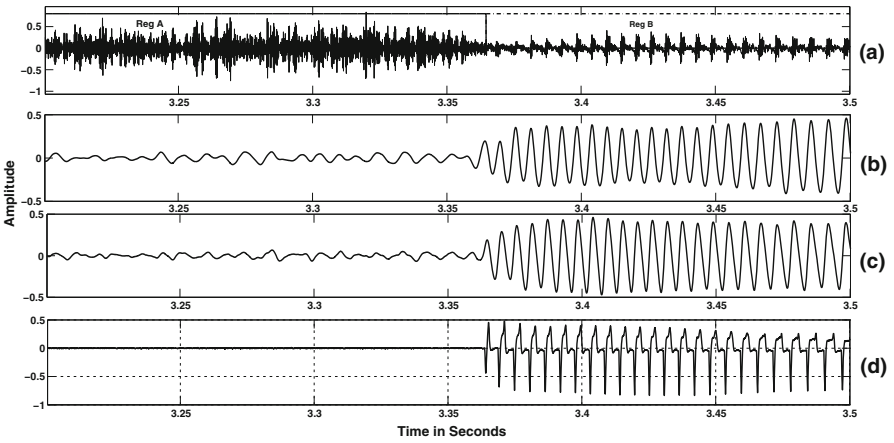

Fig. 13 Robustness of ZFFS and ZBFS for background noise. **a** Speech segment from a noisy background, where RegionA (marked by *solid line*) consists of only background noise and RegionB (marked by *dotted line*) consists of both background noise and foreground speech. **b** Zero frequency filtered output for speech waveform shown in (**a**), **c** zero band filtered output for speech waveform shown in (**a**), **d** differenced EGG which acts as reference epoch locations

can be noisy and speech collected in such environments can have high background noise. The background noise can include speech and non speech like degradations. The desired speaker's speech who is speaking close to sensor (headphone mounted to head and sensor close to mouth) is termed as *foreground speech* and rest of the content are termed as *background degradations*. It is observed that the features such as epoch strength and normalized first order autocorrelation coefficients derived from ZFFS offer discriminatory information between foreground and background regions in speech [5]. Hence, these two features are used to segment *foreground speech* from rest of the *background degradation*. However, these two features depend on the ability of ZFF to extract epoch locations in foreground regions, even in the presence of background noise levels comparable to foreground regions. Even though ZBF allows narrow band of frequencies around 0 Hz to pass through relative to ZFF, this does not impact on the ability of ZBF to extract epoch locations in foreground speech regions. It can be observed from Fig. 13 that both ZFFS and ZBFS are almost similar in their performance, where Fig. 13(a) is a speech segment from a naturally recorded speech in noisy environment and Fig. 13(b) and (c) are ZFFS and ZBFS of a speech segment shown in Fig. 13(a), respectively. It can be observed in Fig. 13(a) that there is presence of background noise through out the speech segment, *RegionA* consists of only background region while *RegionB* consists of both background noise and foreground speech. Also, EGG is recorded in parallel channel to provide the ground truth of epoch locations, where Fig. 13(d) shows the plot of differenced EGG signal that acts as reference.

In order to demonstrate the robustness of the proposed method for *foreground speech* under degraded conditions, speech was collected at different noisy environments along with simultaneous recordings of EGG in parallel channel. Data was collected from 9 males and 4 females speakers. The recording setup had laptop, headphone and EGG

Table 2 Performance evaluation of epoch extraction methods using data collected naturally from different noisy environments

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
DYPSA	82.24	13.22	3.59	0.29
ILPR	89.92	8.54	1.63	0.31
SEDREAMS	73.43	21.29	5.34	0.39
YAGA	91.16	1.24	7.81	0.23
ZBF	92.62	0.43	6.98	0.37
ZFF	91.49	2.31	6.27	0.26

IDR identification rate, *MR* miss rate, *FAR* false alarm rate, *IDA* identification accuracy

electrodes carried to respective locations to record the data from different subjects. Locations included mechanical workshop, home (TV switched on with loud volume), generator room, traffic, dining hall and vehicle noises. The performance was evaluated using EGG as the ground truth and it is given in Table 2. A total of 26,366 epochs were present in the data collected. It can be observed from Table 2 that the performance of ZBF is significantly better in terms of IDR compared to SEDREAMS, DYPSA and ILPR methods and comparable to ZFF and YAGA methods. Furthermore it can be noticed that the performance of ZBF is slightly better to ZFF in terms of IDR. It can be noticed from Eq. (4) that the ZFF output is sensitive to accurate estimation of pitch period from the speech signal. However, it is a challenging task to correctly estimate pitch period in noisy cases and this impacts the results of ZFF in noisy cases. Since, the ZBF is independent of pitch period estimation and this may explain the reason for better IDR in case of ZBF relative to ZFF. However, YAGA, ZFF, and DYPSA are better in terms of IDA compared to other methods.

4.2 Varying Pitch Scenarios

4.2.1 Epoch Extraction in Emotional Data

Epoch extraction in emotional speech data is of significant interest amongst speech synthesis community. It is shown that the excitation source information features derived using epoch extraction methods are significantly different across different emotions. Also, due to large variations in pitch, epoch extraction is a challenging task [8]. In order to demonstrate the robustness of the proposed method for epoch extraction in such emotional data, different emotional speech files containing angry, disgust, fear and happy emotions were chosen from German emotional speech database for evaluation [7]. The database consists of both speech and parallel EGG recordings that is used for reference marking of epochs.

4.2.2 Epoch Extraction in Singing Voice

Epoch extraction in singing voice is challenging due to variations in pitch. In order to evaluate the performance of epoch extraction methods a singing database was created

from 10 different singers consisting of 5 males and 5 females professional singers. All the singers had an average experience of approximately 15 years of practicing Indian Hindustani classical singing and their experiences varied from 5 to 30 years. Each of those singers have sung 2 different songs in order to capture the variations in the pitch. The songs recorded are of different "raag", "taal" and scale, also the lyrics of each of the songs were different and they consists of Hindi and Assamese languages. The singers were made to wear EGG electrodes around the neck along with headphone to capture EGG data along with singing voice, while the songs were recorded in professional studio environment. The portions of the songs were selected that had significant variations in the pitch and all methods were evaluated using 17 min of singing files that consisted of 2,01,808 epochs.

In order to evaluate the performance of ZFF in case of varying pitch scenarios two different versions of ZFF are used in the current study. Firstly, in case of non adaptive ZFF, the average pitch period is measured for the entire speech file that is used as an input parameter for trend removal procedure. However, measuring average pitch for the entire speech file may not be suitable for varying pitch scenarios that can result in spurious positive zero crossings or missing the instants of significant excitation and therefore requires pitch estimation in short term basis. Consequently the updated average pitch ($F0$) is computed using ZFFS for every 20–30 ms of non overlapping frames by picking the maximum peak from STFT of the corresponding frame [8,10]. Furthermore, the updated $F0$ is used as an input to design a low pass filter from which the modified ZFFS signal is obtained and this is called as adaptive ZFFS [8]. It can be observed from Tables 3 and 4 that the performance of ZBF is significantly better than SEDREAMS, ILPR, and DYPSA methods in terms of IDR. However, the performance of ILPR, YAGA, and ZFF are better in terms of IDA. Furthermore, it can be observed from Tables 3 and 4 that the performance of ZBF is better than non adaptive ZFF in terms of IDR while ZFF is having better IDA. The reason for better IDR in case of ZBF compared to non adaptive ZFF may be because the parameters are fixed and they are not dependent on $apriori\ F0$ estimation, while the poor estimation of $F0$ can lead to increase in MR or FAR and this is evident from Tables 3 and 4 that there is an increase in MR. The miss detection are eliminated significantly using adaptive ZFF and as a result there is a significant improvement in terms of IDR. However, the performance of ZBF without $apriori$ pitch estimation is significantly better than SEDREAMS, ILPR and DYPSA for varying pitch scenarios such as emotional and singing voice.

5 Summary and Conclusions

In this paper a bounded input bounded output stable realization of ZFF is proposed and the advantages of using a stable filter is that the method does not require a priori estimation of pitch and furthermore eliminates the necessity of removing trend from the filter's output to obtain ZBFS. The method is validated using three different databases taken from CMU-Arctic for both clean and degraded conditions. It is found that the performance of ZBF is comparable to ZFF. Also, the method is robust for lengthy files recorded for several minutes without any precision related problems associated with the filter output. In order to demonstrate the robustness of ZBF for lengthy files,

Table 3 Performance evaluation of epoch extraction by ZFF and ZBF using angry, disgust, fear and happy emotional speech files from German emotional database

Emotion	Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
	DYPSA	83.32	4.59	12.08	0.54
	ILPR	85.88	10.62	3.50	0.49
	SEDEREAMS	61.19	20.08	18.73	0.51
Angry	YAGA	90.93	4.43	4.64	0.46
	ZBF	91.73	1.49	6.77	0.44
	ZFF (non adaptive)	88.96	0.05	10.97	0.35
	ZFF (adaptive)	92.81	0.00	7.19	0.37
	DYPSA	84.03	7.26	8.71	0.52
	ILPR	89.73	5.22	5.05	0.38
	SEDEREAMS	71.25	16.33	12.42	0.49
Disgust	YAGA	86.72	5.39	7.88	0.43
	ZBF	93.26	3.57	3.17	0.42
	ZFF (non adaptive)	85.51	7.26	8.71	0.34
	ZFF (adaptive)	95.74	1.49	2.77	0.38
	DYPSA	84.61	9.31	6.08	0.55
	ILPR	87.08	8.41	4.51	0.45
	SEDEREAMS	64.70	24.33	9.97	0.58
Fear	YAGA	87.84	6.61	5.55	0.41
	ZBF	94.46	1.18	4.36	0.48
	ZFF (non adaptive)	88.01	3.26	8.73	0.34
	ZFF (Adaptive)	96.20	2.33	1.73	0.44
	DYPSA	85.77	9.48	4.76	0.49
	ILPR	84.03	12.58	3.40	0.51
	SEDEREAMS	70.25	17.10	12.65	0.50
Happy	YAGA	90.01	4.96	5.03	0.43
	ZBF	92.60	3.23	4.16	0.44
	ZFF (non adaptive)	87.43	3.19	9.38	0.33
	ZFF (adaptive)	94.47	1.55	3.98	0.39
	DYPSA	84.30	7.66	7.9	0.52
	ILPR	86.68	9.20	4.00	0.45
	SEDEREAMS	66.84	19.46	13.44	0.52
Average	YAGA	88.87	5.34	5.77	0.43
	ZBF	93.01	2.36	4.61	0.44
	ZFF (non adaptive)	87.47	3.44	9.44	0.34
	ZFF (adaptive)	94.80	1.34	5.66	0.39

IDR identification rate, *MR* miss rate, *FAR* false alarm rate, *IDA* identification accuracy

several speech files of CMU-Arctic database are concatenated to form a lengthy file to which the performance was evaluated. It is found that the performance of ZBF does not

Table 4 Performance evaluation of epoch extraction by ZFF and ZBF using singing database

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
DYPSA	71.90	26.00	2.09	0.26
ILPR	73.56	26.01	0.40	0.18
SEDREAMS	69.95	21.86	8.19	0.31
YAGA	81.50	16.34	2.16	0.23
ZBF	83.08	14.59	2.43	0.29
ZFF (non adaptive)	80.24	14.39	5.36	0.22
ZFF (adaptive)	84.76	13.73	2.51	0.26

IDR identification rate, *MR* miss rate, *FAR* false alarm rate, *IDA* identification accuracy

vary with the length of the files. Furthermore the robustness of ZBF is demonstrated in cases of emotional and singing voices, where the performance is comparable to adaptive ZFF while it is significantly better than SEDREAMS, ILPR and DYPSA for varying pitch scenarios. Future work may focus on improving the performance of ZBF in terms IDA and explore the usefulness of ZBF in different applications.

Acknowledgments This work is part of the ongoing project on the development of Prosodically guided phonetic Engine for Assamese language funded by the Technology Development for Indian Languages (TDIL) Programme initiated by the Department of Electronics & Information Technology (DeitY), Ministry of Communication & Information Technology (MC&IT), Govt. of India under the consortium mode headed by IIIT Hyderabad.

References

1. T.V. Ananthapadmanabha, B. Yegnanarayana, Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Process. **27**, 309–319 (1979)
2. M. Brookes, P.A. Naylor, J. Gudnason, A quantitative assessment of group delay methods for identifying glottal closures in voiced speech. IEEE Trans. Audio Speech Lang. Process. **14**(2), 456–466 (2006)
3. P. Chetana, N. Dhananjaya, S.V. Gangashetty, Analysis of acoustic events in speech signals using bessel series expansion. Springer Circuits Syst. Signal Process. **32**, 2915–2938 (2013)
4. CMU-ARCTIC Speech Synthesis Databases. [Online]. Available: http://festvox.org/cmuarctic/index.html
5. K. T. Deepak, B. D. Sarma, and S. R. M. Prasanna, Foreground speech segmentation using zero frequency filtered signal, in *Interspeech* (2012)
6. T. Drugman and T. Dutoit, Glottal closure and opening instant detection from speech signals, in *Interspeech*, (2009)
7. German Emotional Speech Database. [Online]. Available: http://database.syntheticspeech.de/
8. D. Govind and S. R. M. Prasanna, Epoch extraction in emotional speech, in *SPCOM*, (2012)
9. D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, Neutral to target emotion conversion using source and suprasegmental information, in *Interspeech*, (2011)
10. S. Guruprasad, B. Yegnanarayana, Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals. IEEE Trans. Audio Speech Signal Process. **19**, 1853–1864 (2011)
11. K.S.R. Murthy, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio, Speech Lang. Process. **16**, 1602–1613 (2008)
12. P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE Trans. Audio, Speech Lang. Process. **15**(1), 34–43 (2007)

13. NIST-Speaker Recognition Evaluations. in. [Online]. Available: http://www.nist.gov/itl/iad/mig/sre12.cfm
14. Noisex-92. in. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html
15. A.V. Oppenheim, R.W. Schafer, J.R. Buck, *Discrete-Time Signal Processing* (Prentice-Hall, Upper Saddle River, 1999)
16. A.P. Prathosh, T.V. Ananthapadmanabha, A.G. Ramakrishnan, Epoch extraction based on integrated linear prediction residual using plosion index. IEEE Trans. Audio Speech Lang. Process. **21**, 2471–2480 (2013)
17. K.S. Rao, S.R.M. Prasanna, B. Yegnanarayana, Determination of instants of significant excitation in speech using hilbert envelope and group delay function. IEEE Signal Process. Lett. **14**, 762–765 (2007)
18. R. Smits, B. Yegnanarayana, Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Speech Audio Process. **3**, 352–333 (1995)
19. K.S.S. Srinivas, K. Prahallad, An FIR implementation of zero frequency filtering of speech signals. IEEE Trans. Audio Speech Lang. Process. **20**(9), 2613–2617 (2012)
20. H.W. Strube, Determination of the instant of glottal closure from the speech wave. J. Acoust. Soc. Am. **56**, 1625–1629 (1974)
21. M.R.P. Thomas, P.A. Naylor, The SIGMA algorithm: a glottal activity detector for electroglottographic signals. IEEE Trans. Audio Speech Lang. Process. **17**, 1557–1566 (2009)
22. M.R.P. Thomas, J. Gudnanson, P.A. Naylor, Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. IEEE Trans. Audio Speech Lang. Process. **20**, 82–91 (2012)
23. B. Yegnanarayana, S. R. M. Prasanna, and S. Guruprasad, Study of robusness of zero frequency resonator method for extraction of fundamental frequency, in *ICASSP*, (2011)
24. B. Yegnanarayana, S.V. Gangashetty, Epoch-based analysis of speech signals. Sadhana **36**, 651–697 (2011)
25. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis, in *Eurospeech*, (1999)