INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Robust and Alternative Approach to Zero Frequency Filtering Method for Epoch Extraction

P. Gangamohan and B. Yegnanarayana

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad, India
gangamohan.p@students.iiit.ac.in and yegna@iiit.ac.in

Abstract

During production of voiced speech, there exists impulse-like excitations due to abrupt closure of vocal folds. These impulse-like excitations are often referred as epochs or glottal closure instants (GCIs). The zero frequency filtering (ZFF) method exploits the properties of impulse-like excitation by passing a speech signal through the resonator whose pole pair is located at 0 Hz. As the resonator is unstable, the polynomial growth/decay is observed in the filtered signal, thus requiring a trend removal operation. It is observed that the length of the window for trend removal operation is critical in speech signals where there are more fluctuations in the fundamental frequency (F_0). In this paper, a simple finite impulse response (FIR) implementation is proposed. The FIR filter is designed by placing large number of zeros at $\frac{f_s}{2}$ Hz (f_s represents the sampling frequency), closer to the unit circle, in the z-plane. Experimental results show that the proposed method is robust and computationally less complex when compared to the ZFF method.

Index Terms: Impulse-like excitation, epoch extraction, zero frequency filtering, adaptive zero frequency filtering, zero band filtering

1. Introduction

There are two primary modes of larynx activity, namely, voiced and unvoiced. During voiced mode, the vocal folds vibrate due to tensed larynx, i.e., the process of repeated adduction and abduction takes place. One cycle of vocal fold adduction and abduction is called a glottal cycle. In a glottal cycle, the major impulse-like excitation occurs during vocal fold abduction due to abrupt closure. This impulse-like excitation is referred as the epoch or the glottal closure instant (GCI) [1]. The identification of these epoch locations are useful in many applications like voice source analysis [2], speech pathology [3], prosody modification [4] and emotional speech analysis [2, 5].

The following are the well known approaches for epoch extraction: Algorithm using Hilbert envelope of the linear prediction (LP) residual [6], dynamic programming phase slope algorithm (DYPSA) [7], yet another GCI algorithm (YAGA), algorithm using integrated linear prediction residual [8], zero frequency filtering (ZFF) method [1], and speech event detection using the residual excitation and a mean-based signal (SEDREAMS) method [9]. Most of these techniques [6, 7, 8] use an estimate of the excitation source in the form of LP residual signal, thus relying on source-system separation.

In the ZFF method, the underlying idea is to explore the spectral characteristics of impulse-like excitation. The impulse-like excitation results in discontinuities in the vocal tract system response [1]. These discontinuities are reflected across all frequencies in the spectrum, including at 0 Hz. The vocal tract system characteristics are least effective at 0 Hz. To emphasize the excitation source, and to reduce the effect of the vocal tract system, the speech signal is passed through a 0 Hz resonator. This is followed by trend removal operation. The negative to positive zero crossings in the trend removed signal gives epoch locations. Inspired by the ZFF method, SEDREAMS method explores the analysis by focusing on a mean-based signal. This method uses a mean based signal computed by smoothening the speech signal using Blackman window. In these methods, there is no source-system separation involved. But they require priori knowledge of the pitch period. The information of local pitch period is more critical while processing emotional speech signals, this is due to more fluctuations in the pitch period [10].

In this study, an alternate, computationally less complex and robust ZFF approach is proposed. The filter used in the ZFF method has a pole pair at 0 Hz, on the unit circle, in the z-plane. This unstable filter results an output which has polynomial growth/decay as a function of time. In the proposed method, the zero frequency emphasis is given heavily by designing a filter with large number of zeros at $\frac{f_s}{2}$ Hz (f_s refers to the sampling frequency), in the z-plane. The resultant filter output does not have polynomial trend, thus not requiring trend removal operation. The paper is organized as follows: In Section 2, an overview of ZFF method is presented. In Section 3, we discuss the implementation of proposed method. In Section 4, the performance of proposed method is presented. Finally, Section 5 presents summary and conclusions.

2. An overview of ZFF approach

In the ZFF method, a speech signal is passed through a resonator called zero-frequency resonator. The zero-frequency resonator has a pole pair located at 0 Hz, on the unit circle, in the z-plane. The transfer function of such a resonator is given by

$$H[z] = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}}, \tag{1}$$

where $a_1 = -2$ and $a_2 = 1$. The equivalent operation in the time domain is given by

$$y_1[n] = -2y_1[n-1] + y_1[n-2] + x[n], \tag{2}$$

where $x[n]$ and $y_1[n]$ are the input and output signals, respectively. The input speech signal is passed through a cascade of two such resonators. The resulting output signal $y_2[n]$ has a polynomial growth as shown in Fig. 1(b). The trend removal operation is then applied on the $y_2[n]$ signal, which is given by

$$y_3[n] = y_2[n] - \frac{1}{2M+1} \sum_{p=-M}^{M} y_2[n+p], \tag{3}$$

where $2M + 1$ is the size of the window. The resultant signal $y_3[n]$ is called the ZFF signal. Typically, the window size for

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1172

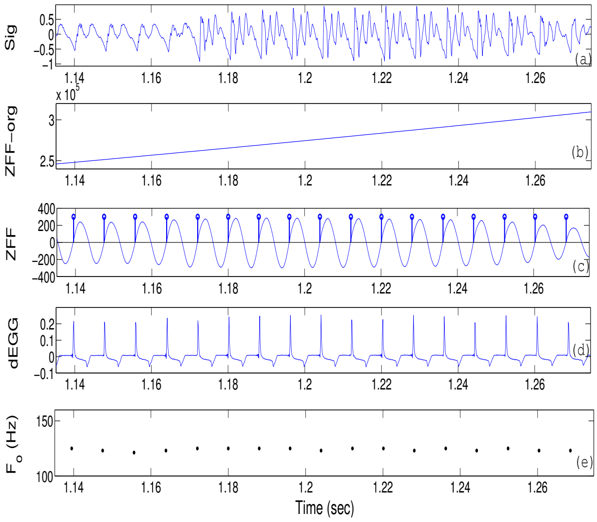

Figure 1: *Illustration of zero frequency filtering (ZFF). (a) Speech signal (Sig). (b) Zero frequency filtered signal before trend removal (ZFF-org). (c) ZFF signal with epoch locations (ZFF). (d) differenced EGG (dEGG) signal. (e) F_0 contour.*

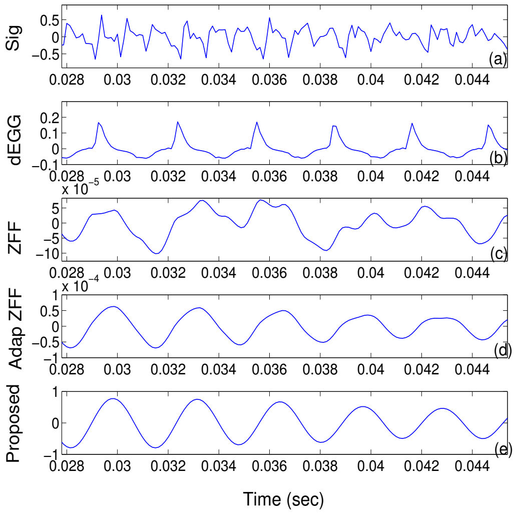

Figure 2: *(a) Segment of a speech signal. (b) Corresponding dEGG signal. (c) Output signal from the original ZFF method. (d) Output signal from the adaptive ZFF method. (e) Output signal from the proposed method (which is discussed in Section 3).*

the trend removal operation is chosen to be approximately about 1.5 times the estimated pitch period. The ZFF signal oscillates at local fundamental frequency (F_0), and the discontinuities due to impulse-like sequence are reflected at the negative to positive zero crossings [1]. The ZFF signal along with the original (neutral) speech signal, differenced electroglottograph (dEGG) signal and F_0 contour are shown in Fig. 1. From the ZFF signal shown in Fig. 1(c), it can be seen that the negative to positive zero crossings are synchronized with the peaks in the dEGG signal which correspond to the GCIs.

In Fig. 1, the ZFF analysis is carried out on the neutral speech signal. Similarly, the ZFF analysis on the angry speech signal is shown in Fig. 2. The speech segment whose local F_0 is much higher than the average F_0 is shown in Fig. 2 (a). This segment is selected for the purpose of illustration of spurious negative to positive zero crossings in the ZFF signal due to local mean pitch period mismatch for trend removal as shown in Fig. 2 (c). To overcome such issues in the ZFF approach, algorithms such as adaptive ZFF method [11], zero band filtering (ZBF) method [10] and finite impulse response (FIR) implementation of ZFF [12] are proposed in the literature.

In the adaptive ZFF method, the speech signal is processed in terms of segments each about 100 ms. This approach gives reasonably accurate epoch locations by reducing the spurious negative to positive zero crossings. The filtered signal using this method is shown in Fig. 2 (d), where there are no spurious zero crossings.

In the ZBF method, the speech signal is passed through a 0 Hz resonator by placing a pole pair inside the unit circle (with r = 0.99), in the z-plane. This filter is stable and output does not have the polynomial trend, but there is a consistent low frequency bias observed in the output signal. A high pass filter with reasonably low cut-off frequency about 80 Hz is used for the bias removal. The performance of ZFF, adaptive ZFF and ZBF methods on emotional speech data is reported in [10]. It is shown that the performance of ZBF method is better than ZFF method, while it still lags behind adaptive ZFF method.

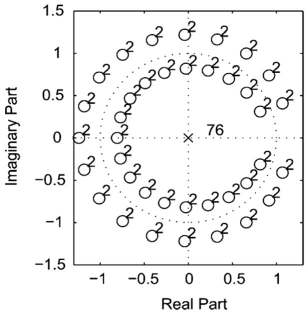

Figure 3: *Pole-zero plot of an FIR implementation of ZFF technique, with order 77. {Figure adopted from [12]}*

In [12], a finite impulse response (FIR) implementation of the ZFF method is proposed. The infinite impulse response (IIR) filter and two FIR trend removal filters used in the ZFF method are derived to a single FIR filter. The pole-zero plot of the FIR implementation is shown in Fig. 3. This filter is a mixed phase system with zeros symmetrically placed across the unit circle, both inside and outside of it. Theoretical derivation of this FIR system suggests that the filter order depends upon the average pitch period, and there are no published results on this implementation. In this paper, an alternative and simple FIR implementation is proposed, which is discussed in the following section. Note that all signal processing computations in this paper are done using a sampling frequency of 8000 Hz.

3. Proposed algorithm

In the proposed method, the zero frequency component is heavily emphasized by placing zeros at $\frac{f_s}{2}$ Hz, with r = 0.99, in the z-plane. The pole-zero plot with 400 zeros at $\frac{f_s}{2}$ Hz in the z-

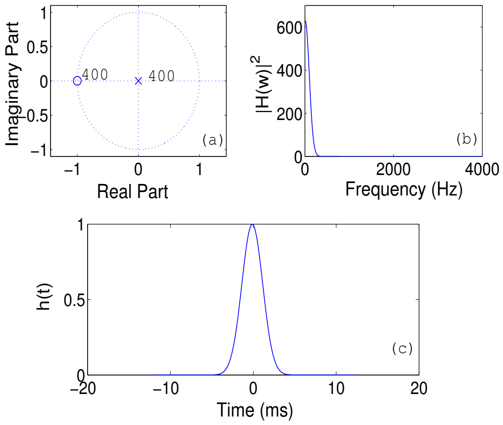

Figure 4: *(a) Pole-zero plot of the proposed FIR implementation. (b) Magnitude response of the filter. (c) Impulse response of the filter.*

plane, its magnitude spectrum and impulse response are shown in Figs 4 (a), (b) and (c), respectively.

A neutral speech signal of a male speaker and the output signal from a FIR filter with 400, 700 and 1000 zeros at $\frac{f_s}{2}$ Hz in the z-plane are shown in Figs 5 (a), (b), (c) and (d), respectively. It is observed that there is no polynomial trend in the output signals, but there are some spurious negative to positive zero crossings in the output signal from a filter with 400 zeros. When the same signal is passed through filters with 700 and 1000 zeros, respectively, there are no spurious zero crossings in the output signal. Also, the negative to positive zero crossings give the epoch locations. As we increase the number of zeros at $\frac{f_s}{2}$ Hz in the z-plane, the impulse response of the system spreads over the function of time. The higher the pitch period, more the spread in the impulse response is required for smoothing the speech signal. From several observations, the optimal number of zeros at $\frac{f_s}{2}$ Hz for the following range of pitch periods: $8-10$ ms, $6-8$ ms, $4-6$ ms and less than 4 ms are 1000, 800, 600 and 400, respectively. The decision of number of zeros and local pitch period is not that critical, as can be seen from Fig 2 (e), where the (angry) speech signal is passed through proposed FIR filter with number of zeros chosen based on its average pitch period.

The impulse response of the filter with 1000 zeros at $\frac{f_s}{2}$ Hz observed to be very low (i.e., close to zero) after a range of -6 ms and +6 ms. The output signal obtained after convolution of the speech signal with the impulse response truncated to 12 ms around reference 0 ms is shown in Fig 5 (e). From this plot, it can be seen that there are no effects due to truncation.

4. Performance evaluation

The evaluations are carried out on two databases, namely, Berlin EMO-DB database [13] and IITT-H Telugu emotion database [5]. From these databases, the utterances corresponding to five emotions (neutral, anger, happiness, fear and sadness) are considered. These databases have signals recorded using microphone and electroglottograph (EGG). The acoustic and EGG signals are time aligned manually to compensate delay among them. The reference (or ground truth) epoch locations are obtained from the dEGG signals by picking a prominent peak in each glottal cycle.

Performance of the proposed algorithm is compared with ZFF and adaptive ZFF methods using the following measures:

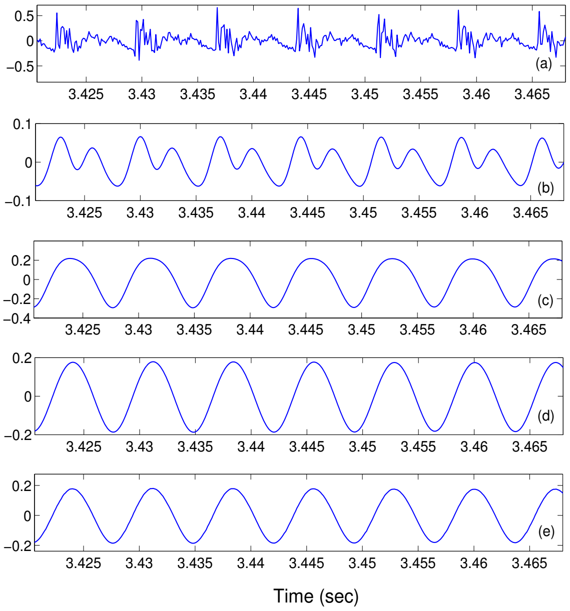

Figure 5: *(a) Segment of a speech signal. (b), (c) and (d) The output signals from the proposed FIR implementation with 400, 700 and 1000 zeros at $\frac{f_s}{2}$, respectively. (e) The output signal obtained after the convolution of the speech with the truncated version of the impulse response of the FIR filter with 1000 zeros.*

- Identification rate (IDR): The percentage of glottal cycles for which only one epoch is detected.
- False alarm rate (FAR): The percentage of glottal cycles for which multiple number of epochs are detected.
- Identification accuracy (IDA): The percentage of glottal cycles with one detected epoch lying in the specified interval. The interval chosen is six samples including reference (ground truth) epoch, i.e., approximately ±0.31 ms. Three samples towards negative time axis and two samples towards positive time axis around the reference epoch are considered.

The results are shown in Table 1. The general observation is that all the methods yeild high performance for neutral speech. The results of the proposed method are much better than ZFF method and matches to that of the adaptive ZFF method. The false alarm rate is observed to be low in the proposed method. Identification accuracy is high in the case of happiness category, and in the other cases it matches to that of the adaptive ZFF method.

For examining the robustness of the proposed method, the performance is evaluated on the speech signals with two additive noise degradations, namely, white noise and babble noise at signal to noise ratio (SNR) levels 20 dB and 10 dB. The noise used is taken from NOISEX database [14]. The averaged IDR and IDA values over all the emotion categories are given in Table 2. From this table, it is clear that the proposed method is robust for these noise degradations.

In terms of memory and computational complexity, the ZFF method requires double precision floating point because of polynomial growth/decay in the output signal from the 0 Hz resonator. Also, the three time trend removal with 1.5 times

Table 1: *Performance evaluation of different methods of epoch extraction for five emotion categories. IDR (in %)−Total identification rate, FAR (in %)−False alarm rate, and IDA (in %)−Identification accuracy within ±0.31 ms.*

Algorithm	Neutral			Anger			Happiness			Fear			Sadness		
	IDR	FAR	IDA	IDR	FAR	IDA	IDR	FAR	IDA	IDR	FAR	IDA	IDR	FAR	IDA
ZFF	94.2	1.3	68.6	85.5	9.8	63.7	86.2	10.6	59.8	90.1	8.3	64.3	88.1	7.8	57.6
adaptive ZFF	96.5	1.0	76.2	94.1	3.3	68.8	93.7	6.1	67.5	93.8	3.6	70.8	92.2	4.9	68.9
Proposed	96.3	0.9	75.1	93.6	3.3	70.4	95.0	3.9	71.4	94.4	2.7	68.7	90.9	4.7	70.5

Table 2: *Average results of different methods for the total data in noisy conditions. IDR (in %)−Total identification rate, FAR (in %)−False alarm rate, and IDA (in %)−Identification accuracy within ±0.31 ms.*

Algorithm	White				Babble			
	(20 dB)		(10 dB)		(20 dB)		(10 dB)	
	IDR	IDA	IDR	IDA	IDR	IDA	IDR	IDA
ZFF	84.5	61.3	74.1	49.5	81.8	54.2	71.8	39.3
adaptive ZFF	91.6	69.8	88.3	53.4	90.3	64.1	83.2	50.7
Proposed	93.3	70.7	86.7	51.7	88.6	67.3	81.4	53.6

of average pitch period contributes to increased computational complexity. In the case of proposed method, single precision floating point is required, and the computational complexity is reduced by 3 times over ZFF method because of no trend removal computation.

5. Summary and conclusions

In this study, a robust and alternative approach of zero frequency (i.e., 0 Hz) filtering is proposed in order to overcome some issues in the zero frequency filtering (ZFF) method. A finite impulse response (FIR) filter is designed by placing large number of zeros at $\frac{fs}{2}$, closer to the unit circle, in the z-plane. This is a stable filter where the output signal does not have polynomial growth/decay, eliminating the necessity of multiple trend removal operations. This reduces the computational complexity by 3 times. Furthermore a priori knowledge of the local pitch period is not that critical in this method.

Robustness of the proposed method is validated on emotional speech data, and on degraded speech data. It is found that the performance of the proposed method outperforms the ZFF method, and is closer and sometimes better than the adaptive ZFF method. This method with very less computational complexity would be effective in low-end processing devices like mobile phone applications.

6. Acknowledgements

The authors would like to thank Tata Consultancy Services (TCS) for partially funding the first author for his PhD programme.

7. References

[1] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[2] P. Alku, "Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[3] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP J. Advances in Signal Processing*, vol. 9, no. 1, pp. 56–75, 2009.

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

[5] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1032–1036.

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

[7] 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, and Language Processing*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[8] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, Dec. 2013.

[9] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, Mar. 2012.

[10] K. T. Deepak and S. R. M. Prasanna, "Epoch extraction using zero band filtering from speech signal," *Circuits, Systems, and Signal Processing*, vol. 34, no. 7, pp. 2309–2333, 2015.

[11] S. A. Thati, K. S. Kumar, and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," *JASA*, vol. 133, no. 5, pp. 3072–3082, 2013.

[12] K. S. S. Srinivas and K. Prahallad, "An FIR implementation of zero frequency filtering of speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2613–2617, Nov 2012.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1517–1520.

[14] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.