EPOCH EXTRACTION FROM SPEECH SIGNALS USING TEMPORAL AND SPECTRAL CUES BY EXPLOITING HARMONIC STRUCTURE OF IMPULSE-LIKE EXCITATIONS

by

Gangamohan P, Suryakanth V Gangashetty

in

International Conference on Acoustics, Speech, and Signal Processing
(*ICASSP-2019*)

Brighton, UK

Report No: IIIT/TR/2019/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
May 2019

EPOCH EXTRACTION FROM SPEECH SIGNALS USING TEMPORAL AND SPECTRAL CUES BY EXPLOITING HARMONIC STRUCTURE OF IMPULSE-LIKE EXCITATIONS

P. Gangamohan and Suryakanth V Gangashetty

Speech Processing Laboratory, Kohli Research Block,
International Institute of Information Technology (IIIT) - Hyderabad, Telangana-500032, India
{*gangamohan.p@research.iiit.ac.in* and *svg@iiit.ac.in*}

ABSTRACT

For voiced speech, significant excitation of the vocal tract system is due to sequence of impulse-like events in the vocal fold vibration. Significant excitations take place around the instants of glottal closure (GCIs), which are often referred as epochs. In literature, most of the epoch extraction algorithms rely on estimation of the excitation source, by attempting to remove the vocal tract system characteristics. The performance of these algorithms depends on the effectiveness of excitation source estimation methods. Algorithms such as zero frequency filtering (ZFF) and speech event detection using the residual excitation and a mean-based signal (SEDREAMS) detect epochs from the smoothened speech signal by exploiting temporal cues. But these algorithms rely on the value of pitch period. The objective of this paper is to develop an algorithm which does not require any apriori information of pitch period, and can be implemented in real-time applications. This algorithm exploits the harmonic property of sequence of impulse-like excitations to obtain the local pitch period, and temporal cues in the filtered signal to obtain epoch locations.

Index Terms— Epoch, instant of glottal closure (GCI), impulse-like excitation, harmonics, pitch period.

1. INTRODUCTION

The repeated process of vocal fold adduction and abduction takes place during the production of voiced speech. In each cycle of vocal fold vibration, the major impulse-like excitation occurs at the glottal closure. The instant of glottal closure (GCI) is often referred to as epoch [1]. Extraction of GCIs from the speech signal is useful in many applications, such as speech enhancement, prosody manipulation, voice conversion, voice modification, and emotion recognition [2, 3, 4, 5]. Most of these applications can be implemented in real-time, where real-time extraction of the GCIs from the speech signal is required.

Epoch extraction algorithms such as dynamic programming phase slope algorithm (DYPSA) [6], yet another GCI algorithm (YAGA) [7], and algorithms proposed in [8, 9] uses linear prediction (LP) residual derived from the LP analysis. Algorithm proposed in [10], exploits the characteristics of glottal flow derivative waveform for extraction of GCIs. These algorithms use an approximate of excitation source in the form of LP residual or glottal flow derivative waveform, thus relying on source-system separation. In general, algorithms based on LP residual depends on the magnitude of discontinuity at epochs, but found to be low for high fundamental frequency (F_0) regions [11]. Also, there are certain issues such as selection of LP order and setting a threshold for an unambiguous detection of GCIs [1, 11].

Algorithms such as zero frequency filtering (ZFF) [1] and speech event detection using the residual excitation and a mean-based signal (SEDREAMS) [12] are based on filtering the speech signal. In the ZFF algorithm, the speech signal is passed through 0 Hz resonator twice, and followed by trend removal operation using local mean subtraction. The window length used for trend removal is observed to be critical in highly varying F_0 regions, especially in the case of emotional speech. To overcome this issue, the modified ZFF (mZFF) algorithm is proposed in [13], where speech signals are processed in segments of duration about 100 ms. In the SEDREAMS algorithm, the speech signal is smoothened using a Blackman window. In a glottal cycle, the GCI lies in the region between valley to peak in the smoothened signal. Magnitude of discontinuity in the LP residual is used to get the accurate location of GCI. However the length of the Blackman window has some effect on performance of GCI extraction. Recently, algorithms based on single frequency filtering (SFF) are proposed in [14, 15, 16]. These algorithms does not require any apriori information of pitch period, but are computationally intensive.

In this paper, we propose a frame level GCI extraction algorithm, which uses two filtering operations on the time domain signal. A filter which has heavily decaying property starting at 0 Hz in the frequency domain, and another filter is a local mean subtracter. These filters are motivated from the studies [1, 17], where time domain characteristics of the filtered signal along with the information of pitch period are used for epoch extraction. In this paper, a detailed spectral analysis of these filters is described. Spectral characteristics are used to extract the GCIs without any apriori information of the pitch period. We consider emotional speech data for evaluations. This is because of performance of the state-of-the-art GCI detection algorithms on emotional speech data are affected due to large variations in the pitch period. The paper is organized as follows: In Section 2, we discuss the steps of the proposed algorithm and illustrate in the cases of synthetic excitation signal and emotional speech segments. In Section 3, results of the proposed algorithm in terms of total identification rate and identification accuracy are presented. Final section gives summary of the paper.

2. PROPOSED ALGORITHM FOR GCI EXTRACTION

The proposed algorithm is based on the following assumptions: 1) The vocal tract system is excited by the sequence of impulse-like excitations. 2) The impulse-like excitation sequence results in harmonic structure in the frequency domain. 3) An impulse-like event in the time domain has relatively higher strength than the strength of the neighboring samples [14].

A filter with heavily decaying spectral response starting at 0 Hz,

and a local mean subtracter are used for epoch extraction. Operation of these filters on a voiced speech segment significantly highlights the first harmonic in the spectral domain, and gives epoch locations. The heavily decaying spectral response filter is designed by placing large number of zeros at $\frac{f_s}{2}$ Hz, on the unit circle, in the z-plane [17]. The transfer function of such a filter with m zeros at $\frac{f_s}{2}$ Hz (where f_s represents the sampling frequency) is given by:

$$H'(z) = (1 + z^{-1})^m. \quad (1)$$

An illustration of $H'(z)$ with $m = 300$ is given in Fig. 1. The pole-zero plot of $H'(z)$, impulse response, and magnitude spectrum are shown in Figs. 1 (a), (b), and (c), respectively. The impulse response ($h'[n]$) of the filter is observed to be very low (or close to zero) after a range of 5 ms to +5 ms. The truncated version ($h[n]$) of the impulse response and its magnitude spectrum are shown in Figs. 1 (d) and (e), respectively. On comparison of Figs. 1 (c) and (e), it is evident that there are no severe effects due to truncation. The local mean subtracter is performed by subtracting average over 10 ms at each sample, which is given by:

$$y[n] = x[n] - \frac{1}{2M+1} \sum_{m=-M}^{M} x[n+m], \quad (2)$$

where $2M + 1$ corresponds to the number of samples in the 10 ms interval. Note that all the signal processing operations in this paper are carried out using a sampling frequency of 8000 Hz.

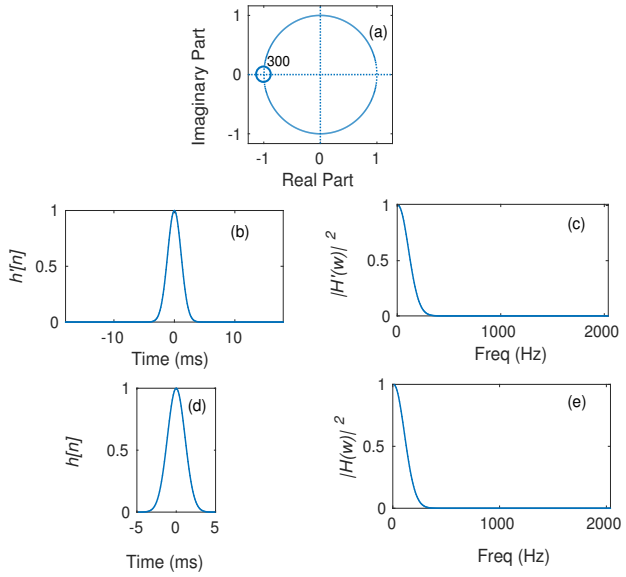

Fig. 1. *(a) Pole-zero plot of the proposed filter. (b) Impulse response ($h'[n]$). (c) Magnitude spectrum ($|H'(\omega)|^2$) of the filter . (d) Truncated impulse response ($h[n]$). (e) Magnitude spectrum ($|H(\omega)|^2$) of the truncated impulse response. Note: The magnitude spectra are amplitude normalized by dividing with respective to the maximum values.*

An illustration of the above filter operations on the synthetic impulse sequence ($s[n]$) is given in Fig. 2. The magnitude spectrum of the periodic impulse sequence has harmonics at regular intervals including at 0 Hz, as shown in Fig. 2 (b). Suppose the synthetic impulse sequence $s[n]$ is convolved with $h[n]$, the output signal $x_h[n]$ ($x_h[n] = s[h] * h[n]$, where $*$ represents convolution) has the magnitude spectrum highlighted at the harmonic 0 Hz, as shown in Fig. 2 (d). To remove the harmonic present at 0 Hz, the signal is passed through the local mean subtracter. The resultant signal $s_1[n]$ has a magnitude response with all the harmonics except at 0 Hz, as shown in Fig. 2 (f). To highlight the first harmonic (located at $F_0 = \frac{1}{T_0}$, where T_0 is the pitch period), the signal $s_1[n]$ is convolved with $h[n]$. The output signal x_{h1} ($x_{h1}[n] = s_1[h] * h[n]$) has the magnitude response highlighted at the first harmonic, as shown in Fig. 2 (h). It is also interesting to note that the valleys in the filtered signal x_{1h} corresponds to the impulse locations. This is due to negative polarity of the impulse sequence.

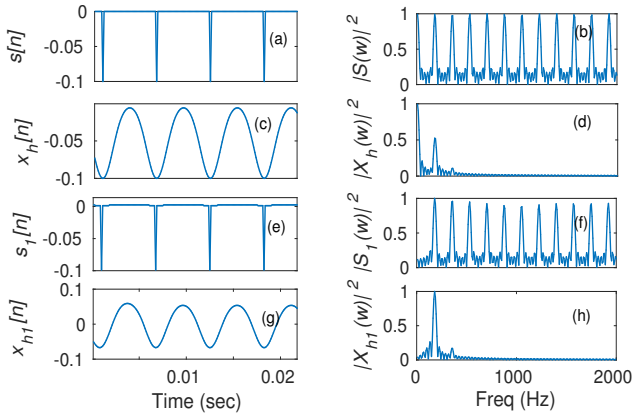

Fig. 2. *(a) Synthetic impulse sequence ($s[n]$). (c) Filtered signal $x_h[n]$ ($x_h[n] = s[n] * h[n]$). (e) Local mean subtracted signal ($s_1[n]$) of the impulse sequence $s[n]$. (g) Filtered signal $x_{h1}[n]$ by ($x_{h1}[n] = s_1[n] * h[n]$). (b), (d), (f), and (h) show the magnitude spectra of (a), (c), (e), and (g), respectively.*

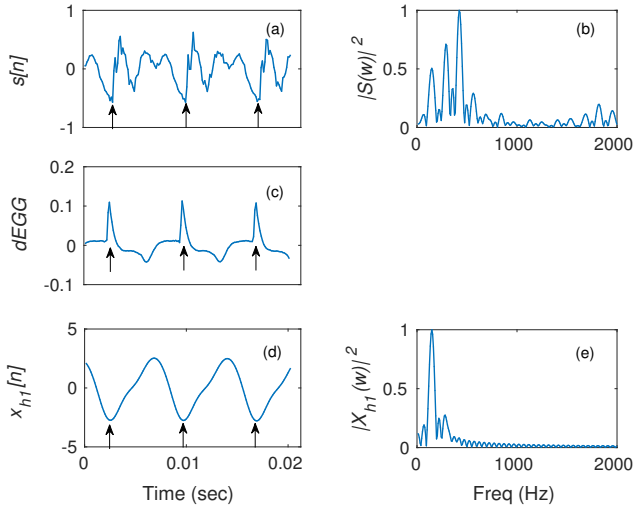

Fig. 3. *(a) Speech signal segment $s[n]$ of a male speaker's neutral speech. (c) dEGG segment corresponding to $s[n]$. (d) Filtered segment $x_{h1}[n]$ using the two filter operations. (b) and (e) Show the magnitude spectra of (a) and (d), respectively.*

Based on the above observations, we attempt to extract the epoch locations in the speech signal. The F_0 of the voiced speech segment is due to the quasi-periodic oscillations caused by the vocal folds vibration. There might be pitch period changes within the analysis frame. The first harmonic moves from F_0 to $F_0 + \delta f$, and the k^{th} harmonic moves from kF_0 to $k(F_0 + \delta f)$ [18]. Due to the larger movement of higher harmonics, only the first few harmonics are evident. The proposed filtering operations heavily emphasize the peak corresponding to F_0.

In the proposed method, speech signals are processed in frames of 32 ms, for every 20 ms. The following are the steps involved:

- Perform local mean subtraction followed by filtering.
- Consider the 20 ms signal from the output filtered signal, i.e., the region between 6 ms and 26 ms.
- Obtain the magnitude spectrum of the selected signal, and locate the first harmonic to compute T_0.
- For each T_0 interval, identify the epoch in the filtered signal by picking the valley. If there are two or more valleys, then consider the valley with highest magnitude.

An illustration of the above steps for a neutral speech segment of a male speaker is given in Fig. 3. The speech segment, corresponding differenced electroglottograph (dEGG) segment, and filtered segment are shown in Figs. 3 (a), (c), and (d), respectively. Figs. 3 (b) and (e) show the magnitude spectra of Figs. 3 (a) and (d), respectively. From Fig. 3 (e), it is evident that the first harmonic is significantly emphasized. Which helps in detecting T_0 unambiguously. On comparison of Figs. 3 (c) and (d), it is clear that the valley with highest magnitude in a cycle in the filtered signal corresponds to the GCI.

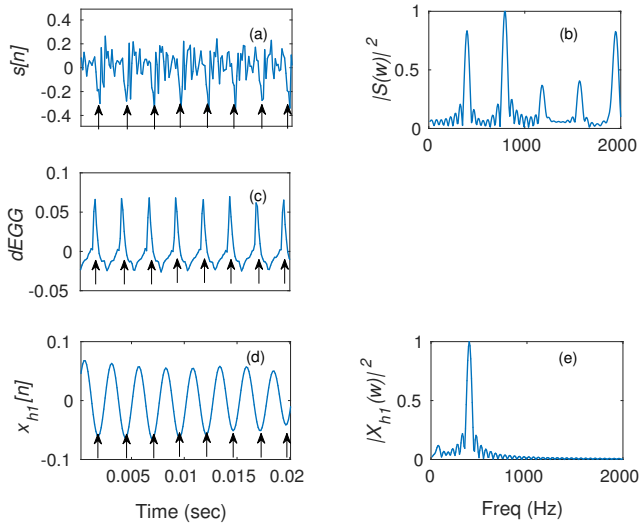

Fig. 4. *(a) Speech signal segment $s[n]$ of a female speaker's angry speech. (c) dEGG segment corresponding to $s[n]$. (d) Filtered segment $x_{h1}[n]$ using the two filter operations. (b) and (e) Show the magnitude spectra of (a) and (d), respectively.*

The proposed algorithm on angry speech segment (with higher F_0) of a female speaker is illustrated in Fig. 4. The speech segment, corresponding dEGG segment, and filtered segment are shown in Figs. 4 (a), (c), and (d), respectively. The magnitude spectra of Figs. 4 (a) and (d) are shown in Figs. 4 (b) and (e), respectively. From these figures, it is evident that the proposed algorithm gives good approximation of GCIs in the case of emotional speech. Similarly, the proposed algorithm on speech segment corresponding to voiced bar /b/ is illustrated in Fig. 5. This figure shows that the proposed algorithm does not depend on the energy of the signal, and gives epochs even in the case of weakly voiced regions.

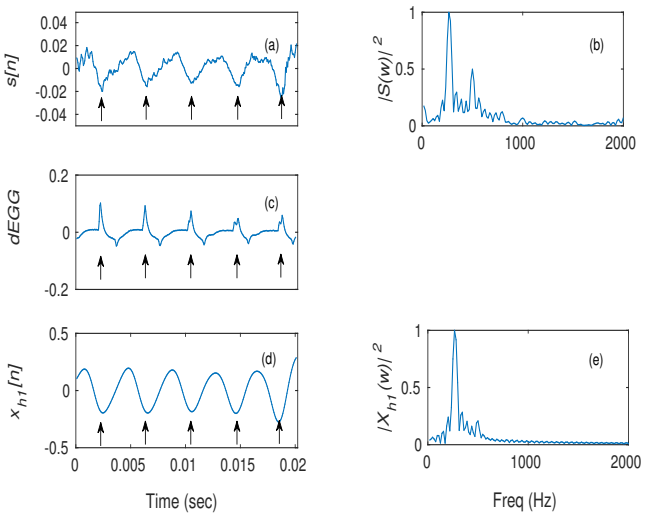

Fig. 5. *(a) Speech signal segment $s[n]$ of a female speaker's neutral speech (voice bar /b/). (c) dEGG segment corresponding to $s[n]$. (d) Filtered segment $x_{h1}[n]$ using the two filter operations. (b) and (e) Show the magnitude spectra of (a) and (d), respectively.*

3. RESULTS AND DISCUSSION

Table 1. *Gross estimation error (GE) (in %) of F_0 detection using the proposed method.*

Emotion category	Gross error (in %)
Neutral	97.2
Anger	94.5
Happiness	95.2
Fear	95.6
Sadness	96.4
Shout	94.9
Surprise	95.5

Detection of GCIs in the analysis frame rely on the accuracy of F_0 estimation. Performance of F_0 estimation is evaluated using the gross estimation error (GE) measure. GE is defined as the percentage of voiced frames with an estimated F_0 value that deviate less than 20% from the reference value [19]. The reference F_0 is computed using the average T_0 value obtained from the corresponding dEGG signal. The average T_0 is computed from the reference epoch locations, which are obtained by picking the prominent peak in each cycle of the dEGG signal. We consider emotional speech data for evaluations. This is because of the performance of the state-of-the-art GCI detection methods on emotional speech data are affected due to large variations in the pitch period. Three databases, namely,

Table 2. *Performance results of ZFF, mZFF, SEDREAMS, and proposed algorithms of epoch extraction for six emotion categories. IDR (in %)−Total identification rate. IDA (in %)−Identification rate to* ± 0.31 *ms.*

	Neutral		Anger		Happiness		Fear		Sadness		Shout		Surprise	
	IDR	IDA	IDR	IDA	IDR	IDA	IDR	IDA	IDR	IDA	IDR	IDA	IDR	IDA
ZFF	93.8	67.3	85.8	63.8	87.5	60.1	90.1	64.3	87.6	58.1	89.9	66.2	83.2	55.2
mZFF	96.3	75.9	93.6	69.4	94.6	67.8	93.7	70.8	92.5	68.9	94.3	72.9	92.6	66.4
SEDREAMS	96.1	76.1	92.7	70.7	93.9	68.2	93.4	71.2	91.7	70.4	94.6	73.7	91.2	69.6
Proposed	95.9	75.5	93.1	70.3	94.4	67.4	94.2	69.9	91.2	69.6	93.8	72.6	93.1	67.5

Berlin EMO-DB [20], IIIT-H Shout corpus [21] and IIIT-H Telugu emotion corpus [5] are considered. From these databases, utterances corresponding to anger, fear, happiness, neutral, sadness, shout, and surprise emotions are used. The average GE (in %) values for different emotion categories are given in Table 1. The results clearly demonstrate that the proposed method gives good estimate of F_0 for the analysis frame, which helps in detecting GCIs unambiguously.

Performance of the proposed algorithm is compared with the ZFF, mZFF, and SEDREAMS algorithms. Two measures, total identification rate (IDR) and identification accuracy to ± 0.31 ms (IDA to ± 0.31 ms), are employed to assess the performance [12]. IDR is a measure of percentage of glottal cycles with only one detected epoch. IDA to ± 0.31 ms is a measure of percentage of glottal cycles with one detected epoch lying in the region ± 0.31 ms around the reference (or ground truth) epoch. Table 2 gives the performance results for IDR (in %) and IDA to ± 0.31 (in %) for the four algorithms ZFF, mZFF, SEDREAMS, and the proposed algorithm. From Table 2, a general observation is that all algorithms give better performance for neutral speech. Among all the algorithms, mZFF algorithm gives better results for IDR, and SEDREAMS gives better performance for IDA to ± 0.31. The performance of the proposed algorithm outperforms ZFF algorithm, and matches to that of the mZFF and SEDREAMS algorithms. The advantages of the proposed algorithm are: 1) It does not require any apriori information of the pitch period. 2) Computationally efficient algorithm, which includes two simple filter operations and fast Fourier transform (FFT) computation, and hence suitable for real-time applications. 3) It does not require any critical thresholds for unambiguous detection of epoch locations. 4) It does not depend on the energy of the signal, and detects epoch locations effectively even in weaker voiced regions.

4. SUMMARY

In this paper, an algorithm for epoch extraction is proposed using temporal and spectral cues. The basic assumption is that the vocal tract system is excited by the sequence of impulse-like excitations. This reflects harmonic characteristics in the spectral domain. The proposed algorithm uses two filtering operations on the time domain signal. Two filters, i.e., a filter with heavily decaying spectral response starting at 0 Hz and a local mean subtracter on the voiced speech segment give unambiguous location of the first harmonic in the spectral domain and epoch locations in the time domain. The performance of the proposed algorithm is compared with other existing algorithms. The identification rate and identification accuracy to ± 0.31 ms are evaluated using three databases, namely, Berlin EMO-DB, IIIT-H Telugu emotion corpus, and IIIT-H shout corpus. Since the proposed algorithm gives better results, and with computationally efficient operations it is suitable for real-time applications

5. 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] 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.

[3] 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.

[4] 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.

[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] 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, 2007.

[7] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[8] 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.

[9] 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, 2013.

[10] A. Koutrouvelis, G. Kafentzis, N. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 316–328, 2016.

[11] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich, and T. Dutoit, "A quantitative comparison of glottal closure instant estimation algorithms on a large variety of singing sounds," in *INTERSPEECH*, Lyon, France, 2013, pp. 1–5.

[12] T. Drugman, M. Thomas, J. Gudnason, P. A. 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, 2012.

[13] T. Sathya Adithya, K. Sudheer Kumar, and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3072–3082, 2013.

[14] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[15] C. M. Vikram and S. R. M. Prasanna, "Epoch extraction from telephone quality speech using single pole filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 624–636, 2017.

[16] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[17] P. Gangamohan and B. Yegnanarayana, "A robust and alternative approach to zero frequency filtering method for epoch extraction," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2297–2300.

[18] T. Bäckström, "Fundamental frequency modelling and estimation," in *Speech Processing Lecture Notes*, Aalto University, 2015.

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

[20] 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.

[21] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3050–3061, 2013.