ANALYSIS OF SINGING VOICE FOR EPOCH EXTRACTION USING ZERO FREQUENCY FILTERING METHOD

Sudarsana Reddy Kadiri and B. Yegnanarayana

Speech and Vision Laboratory
International Institute of Information Technology, Hyderabad, India.
sudarsanareddy.kadiri@research.iiit.ac.in; yegna@iiit.ac.in

ABSTRACT

Epoch is the instant of significant excitation of the vocal tract system during the production of voiced speech. Estimation of epochs or Glottal closure instants (GCIs) is a well studied topic in the speech analysis. From the recent studies on GCI detection from singing voice with state-of-art methods proposed for speech, there exist a clear gap in accuracy between speech and singing voice. This is because of source-filter interaction in singing voice compared to speech. Performance of existing algorithms deteriorates as most of the techniques depends on the ability to model the vocal tract system in order to emphasize the excitation characteristics in the residual. The objective of this paper is to analyze the singing voice for the estimation of epochs by studying the characteristics of the source-filter interaction and the effect of wider range of pitch using the Zero Frequency Filtering (ZFF) method. It is observed that high source-filter interaction can be captured in the form of the impulse-like excitation by passing the signal through three ideal digital resonators having poles at zero frequency, and the effect of wider range of pitch can be controlled by processing short segment (0.4-0.5 sec) signal.

Index Terms— Singing Voice, Excitation Source, Epoch, Glottal Closure Instant, Vocal Tract System, Source-Filter Interaction, Zero Frequency Filtering.

1. INTRODUCTION

The instant of significant excitation of the vocal tract system during the production of voiced speech is referred to as epoch, and it takes place around the glottal closure due to abrupt closing of the vocal folds [1–3]. The importance of epoch extraction and anchoring the analysis around the glottal closure for processing the speech signal has been extensively covered in the recent articles [4, 5]. The field of speech processing has seen a lot of developments in recent years such as creating a variety of techniques for speech analysis, speech modeling/representation and the vocoding techniques for synthesis [6–10]. However, processing techniques for singing voice are not well studied, even though it is closely related to speech signal processing [11–13].

Applying the speech processing techniques for processing of singing voice signals may not be straight forward, even though both are generated from the same production mechanism [11, 12]. This is because many aspects of speech production have been successfully described by a linear source-filter theory [14–17] and in particular Linear Prediction of speech [8, 18] has been the flagship of speech analysis, processing and synthesis. It is recognized that simple linear source-filter theory is not applicable for singing voice [12, 13]. Recently, attempts were made to see the effectiveness of the robust speech signal processing techniques such as pitch extraction algorithms, epoch detection techniques and vocoding techniques for synthesis of singing voice [19–21]. It was found that the usage of the robust speech processing techniques may not be robust for singing voice.

One of the fundamental difference between speech and singing voice is the impact of the source-filter interaction. Since singing voice has more source-filter interaction when compared to speech, it can not be neglected as in most of the speech processing techniques [6, 11, 12, 17, 22]. Apart from the high source-filter interaction, singing voice has wider range of pitch, controlled variations in pitch, variations in phrase duration, prosody, greater dynamic range etc., making the singing voice processing more challenging [12, 13]. In addition to these, the large varieties of singing categories, types and techniques has made it more difficult to generalize the singing voice processing techniques. As a consequence, existing techniques have limited in scope while processing the singing voice signals. For example, in [19], the authors attempted to determine the best method for estimating the Glottal closure instants (GCIs) from the singing voice by evaluating five state-of-art methods of epoch extraction from speech. The choice of the GCI detection algorithm largely depends on the pitch range and singing category. Studies were made to find out the best choice of pitch extraction algorithm and vocoding techniques for singing voice [20, 21]. From studies [19–21], it is clear that there exists noticeable difference in reliability and accuracy of the algorithms.

One of the main weakness of the existing epoch detection techniques is that they depend on the ability to model the vocal tract system in order to emphasize the excitation characteristics in the residual. The objective of the present study is to analyze the source-filter interaction in singing voice using a recently proposed method, namely, the Zero Frequency Filtering (ZFF) [1, 2]. In order to characterize the high source-filter interaction in singing voice, a modified ZFF is proposed. Experimental analysis is carried out for three types of singing voice, and it is found that the proposed method is able to detect the epochs in most of the cases, when compared with the traditional ZFF method used in [1, 3, 19].

The organization of the paper is as follows: Section 2 gives the motivation for the present study. In Section 3.1, analysis of singing voice is carried out using the ZFF method. A modified ZFF is proposed and the analysis for different singing types is presented in Section 3.2. Finally, Section 4 gives a summary and scope for further study.

2. MOTIVATION FOR THE PRESENT STUDY

The motivation for the present study came from the studies [1, 2], where the authors claim that the discontinuities in the excitation signal caused by the sharp closure of the glottis, and they can be approximated by a sequence of impulses of varying amplitudes. The effect of impulse-like excitation is reflected across all the frequencies including the zero frequency. But from the studies in [19], it was found that the ZFF method for singing voice is not as reliable as it is for speech. Unlike other approaches [3, 4, 19], that uses the vocal tract system modeling in order to emphasize the residual, the ZFF method (description of ZFF is given in Sec. 3.1) focuses on filtering the signal at 0 Hz to detect the epoch locations.

The other motivation came from the studies [5, 23–29], where the authors studied the adaptation of ZFF method and its robustness for various types of voices such as laughter, emotion and environments such as distant, telephone, mobile and multi-speaker data. The important characteristics of the ZFF method is that, its instant (abrupt closure of the vocal folds) capturing ability by filtering the signal around 0 Hz.

3. ANALYSIS OF SINGING VOICE USING ZERO FREQUENCY FILTERING METHOD

For the analysis of singing voice, samples from the LYRICS database are considered in this study [30]. The database consists of samples from the 13 trained singers, and the recording sessions took place in a sound-proof booth. It consists of 7 bass-baritones (B1 to B7), 3 countertenors (CT1 to CT3), and 3 sopranos (S1 to S3). Acoustic and electroglottographic signals were recorded simultaneously on the two channels of a DAT recorder. The acoustic signal was recorded using a condenser microphone placed 50 cm from the singers mouth and the electroglottographic signal was recorded using a two-channel electroglottograph. More details of the LYRICS database can be found in [19, 30].

In this study, one of the state-of-art methods of epoch extraction from speech named as Zero Frequency Filtering (ZFF) is used. Here the GCI locations are detected by confining the analysis around a single frequency (0 Hz), i.e., the instant is captured by filtering the signal around 0 Hz.

3.1. Zero Frequency Filtering (ZFF) Method

ZFF method proposed in [1, 2], useful for the extraction of epochs (GCIs), instantaneous fundamental frequency (F_o) and strength of impulse-like excitation [24] by filtering the speech signal through a cascade of two 0 Hz resonators. The advantage of choosing the zero frequency filtering method is that, the characteristics of the time varying vocal tract system will not affect the characteristics of the discontinuities in the output of the resonator.

The following steps are involved to derive the zero-frequency filtered signal:

1. The speech signal $s[n]$ is differenced to remove any unwanted very low frequency components. That is,

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

2. The differenced signal is passed through a cascade of two zero frequency resonators given by,

$$y_o[n] = \sum_{k=1}^{4} a_k y_o[n-k] + x[n] \tag{2}$$

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$. The resulting signal $y_o[n]$ is equivalent to integration (or cumulative sum in the discrete-time domain) of speech signal four times, hence it approximately grows/decays as a polynomial function of time.

3. Using the autocorrelation function, the average pitch period is computed for 30 ms segments of $x[n]$.

4. The trend in $y_o[n]$ is removed by subtracting the local mean computed over the average pitch period at each sample. The resulting signal ($y[n]$) is called as zero frequency filtered signal and is given by,

$$y[n] = y_o[n] - \frac{1}{2N+1} \sum_{i=-N}^{N} y_o[n+i]. \tag{3}$$

where $2N+1$ corresponds to the number of samples in the window used for trend removal.

The instants of negative-to-positive zero crossings (NPZCs) correspond to the significant excitation *epochs or Glottal Closure Instants (GCIs)* by considering the positive polarity of the signal [1, 2, 31]. However, if the speech signal is reversed in polarity, then the signal has to be negated before the epoch extraction [31]. To illustrate this, a segment of speech along with the simultaneously recorded EGG signal from the CMU arctic database is used [1]. Fig. 1 shows the voiced speech segment, ZFF signal along with GCIs marked by arrows and reference differenced EGG (dEGG) signal. It can be seen that, there is a close agreement between the locations of the strong negative peaks of the dEGG signal and the instants of NPZCs derived from the ZFF signal.

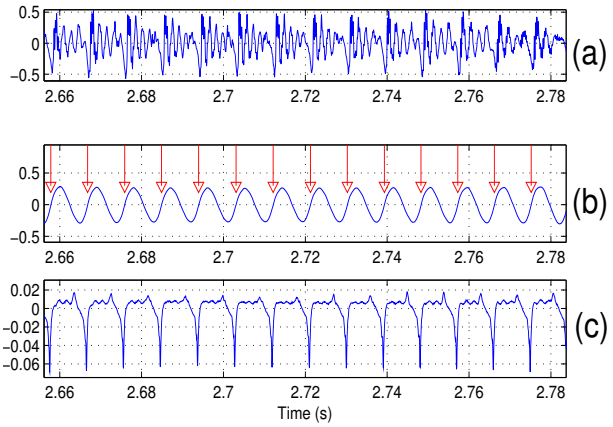

Fig. 1. (a) Segment of a speech signal, (b) Zero-frequency filtered (ZFF) signal (epoch locations marked by arrows), and (c) differenced EGG signal.

Similarly, the analysis for three types of singing voice samples (bass-baritones, countertenors, and sopranos) are performed and they are shown in Figs. 2, 3 and 4, respectively. For the purpose of illustration, we selected the singing voice samples for which the ZFF method failed to capture the impulse-like excitations. In all the figures, (a) is the segment of a singing voice, (b) is the zero-frequency filtered signal (epoch locations are marked by arrows), and (c) is the differenced EGG signal (dEGG) as reference. From Figs. 2, 3 and 4, it is noted that ZFF signal is not able to capture the major impulse-like excitations, unlike the case of speech signal shown in Fig. 1. From Fig. 2, it can be seen that even though the ZFF

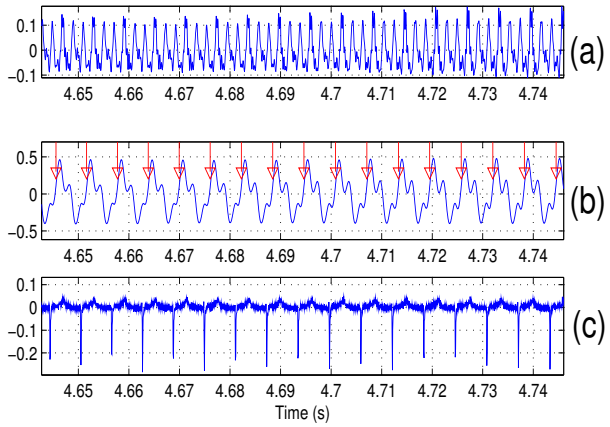

Fig. 2. (a) Segment of a Baritone Singing Voice, (b) Zero-frequency filtered (ZFF) signal (epoch locations marked by arrows), and (c) differenced EGG signal.

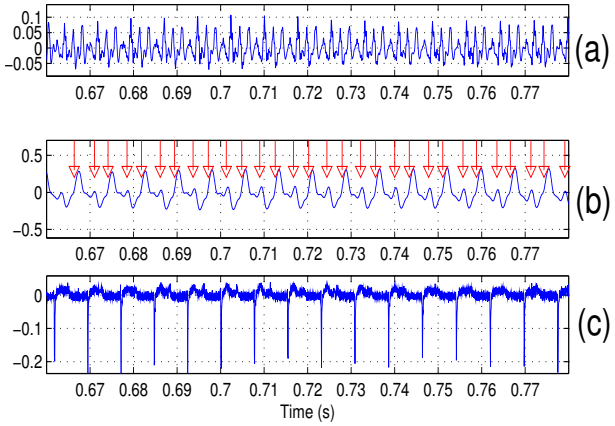

Fig. 3. (a) Segment of a Countertenor Singing Voice, (b) Zero-frequency filtered (ZFF) signal (epoch locations marked by arrows), and (c) differenced EGG signal.

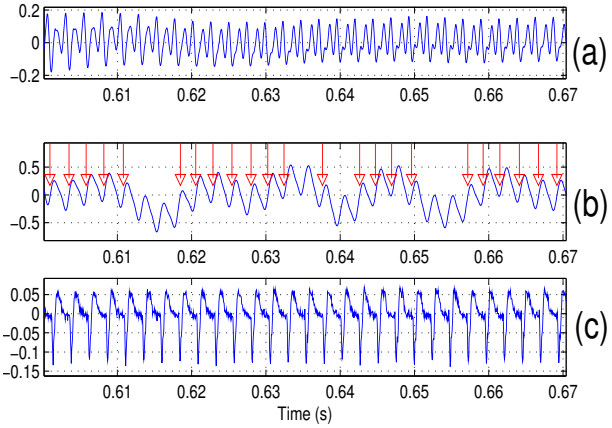

Fig. 4. (a) Segment of a Soprano Singing Voice, (b) Zero-frequency filtered (ZFF) signal (epoch locations marked by arrows), and (c) differenced EGG signal.

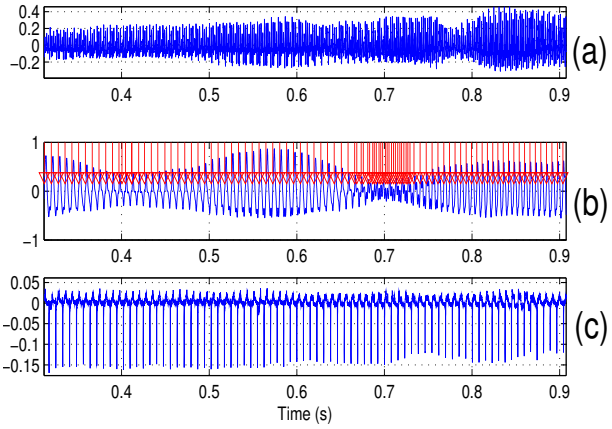

Fig. 5. (a) Segment of a Baritone Singing Voice, (b) Zero-frequency filtered (ZFF) signal (epoch locations marked by arrows), and (c) differenced EGG signal.

signal is giving the reasonable epoch locations, it is not capturing the impulse-like excitations properly when compared to Fig. 1. It is evident from the Fig. 3 also, where the ZFF signal has more number of NPZCs, and hence it is detecting more false GCIs. Due to drift in the ZFF signal (as in Fig. 4), some of the GCIs are missing, and it might be because of the improper trend removal operation.

From Fig. 1, we can mark the epoch locations even without the reference dEGG signal. This is because, within each glottal cycle the excitation of the vocal tract system is impulse-like around the GCI, and it corresponds to the high SNR region due to strong excitation, and also due to decay of the resonances of vocal tract system within each cycle [5, 32]. But it is not the case in Figs. 2, 3 and 4. We can interpret this behavior due to impact of filter interaction in singing voice.

The problem with the singing voice is due to wide range in the controlled variations of pitch. The effect can be seen from the ZFF signal shown in Fig. 5(b) around 0.7 sec. This is because, the ZFF method depends on the average pitch period for trend removal. To overcome this problem, a method is proposed, where short segments (0.4 or 0.5 sec) of the signal are used instead of the total length of the signal (phrase/utterance) to detect the epoch locations. In this case the pitch period can be estimated for each segment, and hence the resulting ZFF signal can capture the impulse-like excitation after the trend removal operation. Also, this way of processing the signal seems to be more realistic in many situations. This is also the case for most of the existing epoch detection algorithms [19] as they depend on the average pitch period.

A modified version of ZFF is proposed to capture the impulse-like excitation that is present in the singing voice, and it is described in Sec. 3.2.

3.2. Modified ZFF Method for Singing Voice

In this section, we propose a modified version of ZFF method to capture the major impulse-like excitation. The modified version has similar steps that are described in the previous section, except that the processing of the signal is now on short segments, and the use of cascade of three ideal digital resonators having poles at 0 Hz. The output of the resonators is given by

$$y_o[n] = \sum_{k=1}^{6} a_k y_o[n-k] + x[n] \tag{4}$$

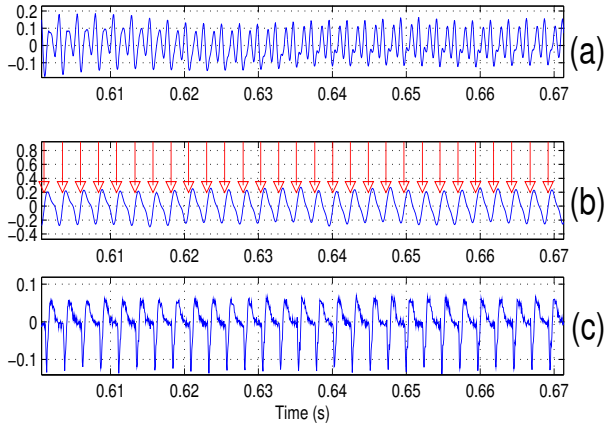

Fig. 6. (a) Segment of a Soprano Singing Voice, (b) Modified zero-frequency filtered signal (epoch locations marked by arrows), and (c) differenced EGG signal.

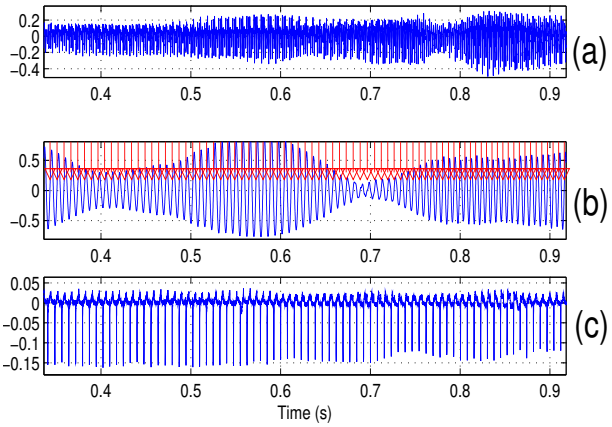

Fig. 7. (a) Segment of a Baritone Singing Voice, (b) Modified zero-frequency filtered signal (epoch locations marked by arrows), and (c) differenced EGG signal.

Table 1. Performance comparison of epoch detection methods on subset of LYRICS database. IDR - Identification rate, MR - Miss rate, FAR - False alarm rate, IDA - Identification accuracy. The first three of the above are collectively called the reliability measures and the other is called the accuracy measure [1, 4].

Method	IDR (%)	MR (%)	FAR (%)	IDA (ms)
ZFF	82.16	1.26	16.58	0.59
Proposed	93.46	1.18	5.36	0.42

where $a_1 = +6$, $a_2 = -15$, $a_3 = +20$, $a_4 = -15$, $a_5 = +6$, $a_6 = -1$.

The trend removal operation is repeated five times in order to get the modified ZFF signal. The remaining steps are same as the ZFF method. It is to be noted that, the passage of signal through cascade of three ZFRs may not be always necessary for epoch detection for all types of singing voices. For some types of singing voice even the traditional ZFF method will give proper epoch locations [3]. But the modified ZFF method gives the impulse-like excitation sequence even for the case of high source-filter interaction in the signals. The output of modified ZFF for segments of singing voice samples given in Figs. 4(b) and 5(b) are shown in Figs. 6(b) and 7(b), respectively. The output of the modified ZFF signals are in close agreement to the reference dEGG signals. Performance of the proposed method along with the traditional ZFF method is given in Table 1. From this, it is observed that the modified ZFF method is able to detect the epoch locations in most of the cases, compared with the traditional ZFF method. Further analysis is required for epoch detection by analyzing other singing types and categories, and also by extending these studies to large databases.

It is worth mentioning that, the analysis carried out in this paper is to understand the impact of source-filter interaction and wider pitch ranges of singing voice. From the preliminary results, it is noted that the proposed modified ZFF method is working consistently for the singing voice compared with the traditional ZFF method.

4. SUMMARY

In this paper, the impact of source-filter interaction and effect of wider pitch range of singing voice were analyzed using ZFF method for the extraction of epochs. From the analysis using ZFF method, it was observed that there exists a high source-filter interaction in various types of singing voice. The effect of wider pitch range on ZFF output was studied, a method was proposed where by processing the short segment of the signal (for e.g, 0.4 or 0.5 sec) instead of total length (phrase/utterance), as the the average pitch period varies rapidly over short segments. This way of processing the signal is closer to realistic situations. A modified version of ZFF method was proposed for epoch extraction by passing the signal through three zero frequency resonators. From the experiments, it was observed that the proposed ZFF method was able to capture the impulse-like excitations (epochs) in most of the cases compared with the traditional ZFF method. The focus in this paper was on the analysis of singing voice for epoch detection. Since the proposed method provides accurate locations of epochs, the results may also be useful for pitch extraction from singing voice. Also, there is scope for understanding the effect of subglottal resonances in singing voice.

5. ACKNOWLEDGEMENTS

The first author would like to thank Dr. Dhananjaya N Gowda, who gave the initial motivation to pursue this work. The authors would like to thank Nathalie Henrich and Onur Babacan for lending the LYRICS database.

6. REFERENCES

[1] K. Sri Rama Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[2] B. Yegnanarayana and K. Sri Rama Murty, "Event-based instantaneous fundamental frequency estimation from speech

signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, May 2009.

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

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

[5] B. Yegnanarayana and Suryakanth V Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no. 5, pp. 651–697, 2011.

[6] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, 2014.

[7] Paavo 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.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio Speech and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[9] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 2, pp. 1039–1064, 2009.

[10] Qiong Hu, Korin Richmond, Junichi Yamagishi, and Javier Latorre, "An experimental comparison of multiple vocoder types," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 155–160.

[11] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. 2733–2749, 2008.

[12] Malte Kob Nathalie Henrich, Hanspeter Herzel, David Howard, Isao Tokuda, and Joe Wolfe, "Analysing and understanding the singing voice : Recent progress and open questions," *Current bioinformatics*, vol. 6, no. 3, pp. 362–374, 2011.

[13] Johan Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, pp. 82–91, 1977.

[14] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, Paris, 2nd edition, 1970.

[15] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[16] G. Fant, "Some problems in voice source analysis," *Speech Communication*, vol. 13, pp. 7–22, 1993.

[17] Flanagan J L, *Speech Analysis, Synthesis and Perception*, Springer, Newyork, 1972.

[18] J. Makhoul, "Linear prediction: A tutorial review," vol. 63, pp. 561–580, Apr. 1975.

[19] Onur Babacan, Thomas Drugman, Nicolas D'Alessandro, Nathalie Henrich, and Thierry Dutoit, "A quantitative comparison of glottal closure instant estimation algorithms on a large variety of singing sounds," in *INTERSPEECH*, 2013, pp. 1702–1706.

[20] Onur Babacan, Thomas Drugman, Tuomo Raitio, Daniel Erro, and Thierry Dutoit, "Parametric representation for singing voice synthesis: A comparative evaluation," in *ICASSP*, 2014.

[21] Onur Babacan, Thomas Drugman, Nicolas D'Alessandro, Nathalie Henrich, and Thierry Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *ICASSP*, 2013, pp. 7815–7819.

[22] I.R. Titze, T. Riede, and P. Popolo, "Nonlinear source-filter coupling in phonation: Vocal exercises," *J. Acoust. Soc. Am.*, vol. 123, pp. 1902–1915, 2008.

[23] B. Yegnanarayana, S. R M Prasanna, and S. Guruprasad, "Study of robustness of zero frequency resonator method for extraction of fundamental frequency," in *ICASSP*, May 2011, pp. 5392–5395.

[24] K. Sri Rama Murty, B. Yegnanarayana, and M. Anand Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Letters*, vol. 16, no. 6, pp. 469–472, June 2009.

[25] D. Govind, S. R. Mahadeva Prasanna, and Debadatta Pati, "Epoch extraction in high pass filtered speech using hilbert envelope," in *INTERSPEECH*, 2011, pp. 1977–1980.

[26] G. Seshadri and B. Yegnanarayana, "Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 1853–1864, 2011.

[27] K. Sudheer Kumar, Sri Harish Reddy Mallidi, K. Sri Rama Murty, and B. Yegnanarayana, "Analysis of laugh signals for detecting in continuous speech," in *INTERSPEECH*, 2009, pp. 1591–1594.

[28] B. Yegnanarayana and S. R. Mahadeva Prasanna, "Analysis of instantaneous f0 contours from two speakers mixed signal using zero frequency filtering," in *ICASSP*, 2010, pp. 5074–5077.

[29] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Process. Letters*, vol. 17, no. 3, pp. 273–276, March 2010.

[30] N. Henrich, C. dAlessandro, M. Castellengo, and B. Doval, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *J. Acoust. Soc. Am.*, vol. 117, no. 3, pp. 1417–1430, 2005.

[31] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *IEEE Sig. Pro. Letters*, vol. 20, no. 4, pp. 387–390, 2013.

[32] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 1, pp. 24–33, Feb 1977.