INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Robust Non-Parametric and Filtering Based Approach for Glottal Closure Instant Detection

Pradeep Rengaswamy, Gurunath Reddy M,
K. Sreenivasa Rao, Pallab Dasgupta

Indian Institute of Technology Kharagpur, WestBengal, India.
rpradeep@iitkgp.ac.in, mgurunathreddy@sit.iitkgp.ernet.in,
ksrao@sit.iitkgp.ernet.in, pallab@cse.iitkgp.ernet.in

Abstract

In this paper, a novel non-parametric based glottal closure instant (GCI) detection method after filtering the speech signal through a pulse shaping filter is proposed. The pulse shaping filter essentially de-emphasises the vocal tract resonances by emphasising the frequency components containing the pitch information. The filtered signal is subjected to non-linear processing to emphasise the GCI locations. The GCI locations are finally obtained by a non-parametric histograms based approach in the detected voiced regions from the filtered speech signal. The proposed method is compared with the two state-of-the-art epoch extraction methods : Zero frequency filtering (ZFF) and SEDREAMS (both of which requires upfront knowledge of average pitch period). The performance of the method is evaluated on the complete CMU-ARCTIC dataset consisting of both speech and Electroglottograph (EGG) signals. The robustness of the proposed method to the additive white noise is evaluated with several degradation levels. The experimental results showed that the proposed method is indeed immune to noise and the obtained results are comparably better than the two state-of-the-art methods.

Index Terms: GCI locations, pulse shaping filter, pitch, histogram

1. Introduction

The major source of excitation to the time-varying vocal tract system is impulse like excitation. The impulsive excitation is due to the glottal activity during the production of voiced speech [1]. The impulsive excitation to the time varying vocal-tract system is manifested as abrupt discontinuity in the produced speech. The discontinuity due to impulsive excitation in the speech signal can be observed in the linear prediction residual (LPR) as either positive or negative peaks [2]. The location of peaks in the LPR roughly corresponds to the GCI of the glottal activity [3] and they are also known as the instants of significant excitation or epoch. The accurate detection of GCIs plays a significant role in most of the speech tasks such as extracting the pitch contour of the quasi-periodic speech signal [4], pitch synchronous analysis of speech [5, 6], epoch based prosody modification [7], melody extraction from vocal polyphonic music signals [8], glottal flow estimation [9], speech synthesis [10], voice source modelling for parametric speech synthesis [11] and so on. In literature, we can find GCI detection methods mostly based on either smoothing (filtering) or computing the energy envelope of the signal prior to GCI detection. The GCI detection methods includes : Hilbert-Envelope [12], ZFF [1], SEDREAMS [13], LOMA [14], YAGA [15] and DYPSA [16]. For more detailed review on GCI detection methods refer [17].

In this work, speech is filtered using Raised Cosine Filter (RCF). The RCF is a pulse shaping low-pass filter widely used in digital communication for minimizing the inter symbol interference. Initially, the RCF filtered speech signal is thresholded to detect the voiced/unvoiced regions. In each identified voiced regions, the peaks corresponding to the GCIs are emphasised by non-linear filtering. Followed by a novel average epoch interval detection method based on the histogram and GCI detection based on peak picking approach is proposed. The source code of the proposed method is available at *https://github.com/Pradeepiit/GCIDetection.git*

2. Proposed Method

The block diagram of the proposed GCI detection method is illustrated in Figure 1. Initially the vocal tract resonances are suppressed by passing the speech signal through RCF, followed by voiced and unvoiced classification, candidate epoch interval (difference between successive GCIs) detection of each voiced region based on histograms, finally GCIs are obtained by peak picking. The implementation details of each block is discussed in the following subsections.

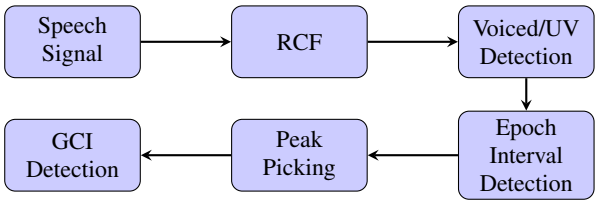

Figure 1: *Block diagram of the proposed GCI detection method.*

2.1. Pre-processing

Initially, the speech signal is passed through a DC removal filter to suppress the slowly varying frequncy components near the $0\ Hz$. The filter is essentially a single pole-zero filter and its transfer function is given by

$$H(z) = \frac{1 - z^{-1}}{1 - \alpha z^{-1}} \tag{1}$$

The discrete time equivalent of the above transfer function is given by

$$x[n] = \alpha x[n-1] + s[n] - s[n-1] \tag{2}$$

where $s[n]$ is the input speech signal, $x[n]$ is the filtered signal. An optimal value of 0.98 is chosen for α, decides the sharpness of the attenuation of slowly varying frequency components near $0\ Hz$.

The RCF [18] is a low-pass, truncated finite impulse response filter with frequency response given by

http://dx.doi.org/10.21437/Interspeech.2016-369

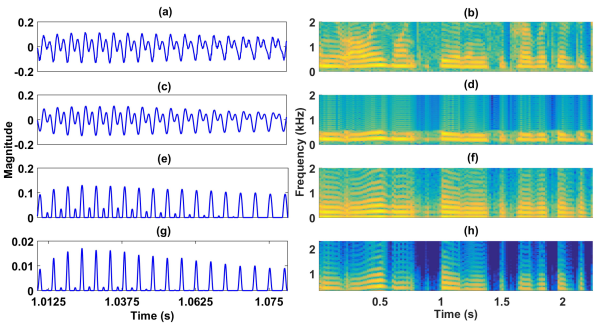

Figure 2: *Illustration of RC filtering of speech signal followed by non-linear processing for emphasizing the epoch locations. A segment of speech signal, RC filtered signal, positive clipped and inverted signal, and resulting squared signal is shown in (a), (c), (e) and (g) respectively. The corresponding spectrogram of the time domain signals are shown in (b), (d), (f) and (h) respectively.*

$$H(f) = \begin{cases} T & , |f| \leq \frac{1-\beta}{2T} \\ \frac{T}{2}\left[1+\cos\left(z\right)\right] & , \frac{1-\beta}{2T} \leq |f| \leq \frac{1+\beta}{2T} \\ 0 & , \text{otherwise} \end{cases}$$

where

$$z = \frac{\pi T}{\beta}\left[|f| - \frac{1-\beta}{2T}\right] \tag{3}$$

The discrete time impulse response is given by

$$h[n] = \frac{\sin \pi n/T}{\pi n/T} \frac{\cos(\pi\beta n/T)}{1-(4\beta^2 n^2/T^2)} \tag{4}$$

Where $T = 1/f_c$, f_c is the pass-band frequency edge and β is the pass-band roll-off factor. As β increases, the sharpness of pass-band edge decreases and hence the frequency content is passed beyond the desired band (an optimal value of 0.25 is chosen for β). For a realizable RCF, the theoretical bandwidth of the filter (B) is given by $B = (1+\beta)f_c$ [19] where, f_c is chosen as 250 Hz which results in a pass-bandwidth of approximately 400 Hz. A slightly higher pass-band frequency than the fundamental frequency range ($85-255$ Hz) of normal male and female speakers is chosen to include the pitch information of other type of signals such as emotional speech and para-linguistic (laughter). The filtered signal $y[n]$ is obtained as the convolution of the signals $x[n]$ and $h[n]$ given by

$$y[n] = x[n] * h[n] \tag{5}$$

where $*$ is the convolution operator.

The output of filtered signal $y[n]$ essentially contains the information about the GCI as predominant peak within a pitch period. A segment of a speech signal, output of the RC filtered speech signal and the corresponding spectrograms are shown in Figure 2(a), (c), (b) and (d) respectively. From Figure 2(c) and (d) we can observe that RCF indeed preserved the discontinuities due to impulsive excitation (GCI locations) as predominant peaks in the filtered signal by removing other frequency contents.

2.2. Voiced/Unvoiced Detection

Voiced and unvoiced (V/UV) segments refers to the glottal and non-glottal activity regions in the speech signal. The V/UV regions in the RCF filtered signal is obtained by thresholding the signal with a factor of $\delta = \frac{1}{12}^{th}$ of the maximum peak to retain even the lower excitation regions in the smoothed signal. The value for δ is obtained by experimental analysis of the dataset.

2.3. Candidate Epoch Interval Detection

The locations corresponding to the significant excitation are manifested as a strong discontinuity (peak or valley) within a pitch period of the voiced signal. The peaks corresponds to the GCIs can be observed as strong peaks within a pitch period. This can be observed in either the negative or positive half of the voiced regions in $y[n]$ along with other significantly comparable peaks. In order to unambiguously detect the locations corresponding to GCI, the peaks corresponding to the epoch locations needs to be emphasized while suppressing the other comparable magnitude peaks. Hence, the smoothed signal is subjected to a non-linear filtering as shown in the block diagram in Figure 3. The significance of each block is explained in the following subsections.

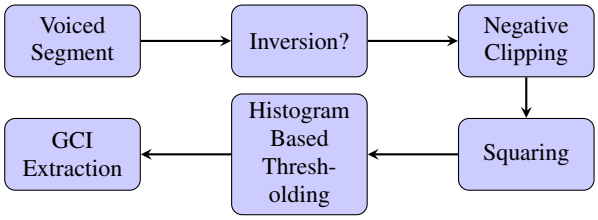

Figure 3: *Block diagram illustration of the non-linear filtering for GCI detection.*

2.3.1. Inversion for Unambiguous Significant Excitation Detection

Instants of significant excitation is shown as either a prominent positive peak or negative valley within a pitch period of the voiced segment. The choice of positive or negative portion of the voiced segment for GCI detection depends on the magnitude deviation between two successive peaks or valleys and the number of peaks or valleys within the pitch period. The decision for clipping and inversion is made based on the following criterion for reliably detecting the GCI locations. The significant peaks in the positive ($\boldsymbol{NV}$) and negative ($\boldsymbol{IV}$) segment of the voiced segment is identified by thresholding $\frac{1}{6}^{th}$ of the maximum peak. This will capture the dominant peaks and other significant peaks. The set with minimum number of peaks is considered for further analysis since, it contains mostly the peaks corresponding to GCI locations. A voiced segment which requires inversion is shown in Figure 4. A segment of voiced speech, filtered signal and the inverted signal is shown in Figure 4(a), 4(b) and 4(c) respectively. From Figure 4(c) we can observe that the prominent peaks corresponding to significant excitation is shown clearly in the positive portion of the signal without any ambiguity.

2.3.2. Negative Clipping

The positive half of the signal containing the significant peaks are retained and the remaining half is clipped off. The clipped signal essentially retains the harmonic content in the signal with an added DC component in the signal which can be observed in Figure 2(e) and 2(f).

2.3.3. Squaring

The clipped signal is further squared to enhances the prominent peaks and to suppress the other peaks within a voiced segment. This has the effect of enhancing the strong excitation regions and suppressing the weak excitation region which is as shown in Figure 2(g) and 2(h) respectively. The negative clipping and squaring leads to non-linearity in the filtered signal.

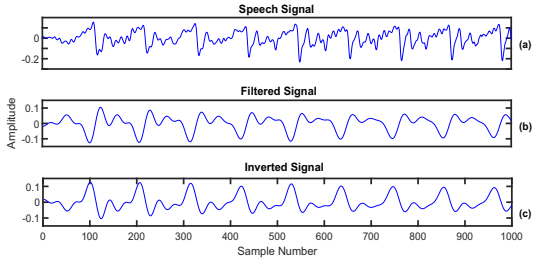

Figure 4: *Illustration of inversion process (a)Voiced speech segment, (b) RC filtered signal and (c) inverted RC filtered signal.*

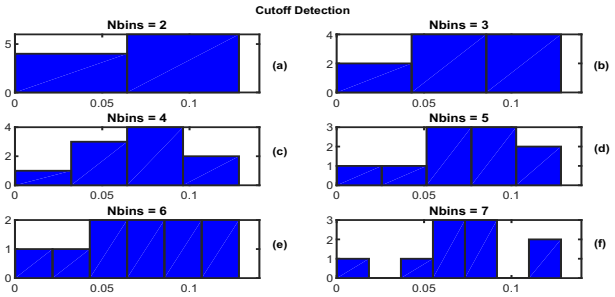

Figure 5: *Illustration of the histogram based segregation of candidate peaks corresponding to GCIs and other peaks. Figures (a),(b),(c),(d),(e), and (f) represent the histograms of the peaks for varying bins of 2,3,4,5,6 and 7 respectively.*

2.3.4. Average Epoch Interval Detection

The non-linear filtered signal contains dominant peaks corresponding to GCIs and other comparable peaks in the weakly voiced regions. An appropriate threshold needs to be defined to discriminate the dominant peaks to other comparable peaks. Hence, a small segment of signal about 50 ms is chosen around the maximum peak in the voiced segment for detecting the average epoch interval. All local peaks of the segment are distributed across a histogram with increasing number of bins. This is performed since, the exact number of bins to segregate the primary peaks is not known in prior. The number of bins are increased linearly with a unit step size starting from two bins until an empty bin is created. An empty bin essentially signifies the peaks are well distributed across bins and a threshold between dominant peaks and other peaks can be obtained. The process of histogram peak distribution is shown in Figure 5.

In Figure 5, the abscissa represents the magnitude of peaks and the ordinate represents the peaks count. From Figure 5(f) we can observe that the dominant peaks are segregated to the right of the histogram and other peaks to the left. A threshold for detecting the average epoch interval (by peak picking and finding the average peak interval) is searched as the bin edge traversing the histogram from right to left as shown in Figure 6(a). The average epoch interval is finalized for the bin edge whose epoch interval deviation is less than 5%. Figure 6(b), 6(c), 6(d), and 6(e) shows the epochs detected for the first four bin edges from the right of the histogram of Figure 6(a). In Figure 6(e), average epoch interval deviates more than 5%, so the average epoch interval will be calculated from Figure 6(d).

2.4. Peak Picking for GCI Detection

The distance between two GCI locations (epoch interval) does not vary drastically within a voiced segment. The average

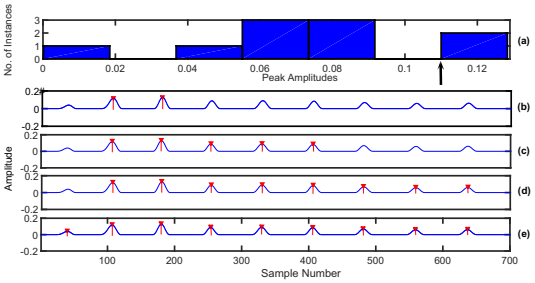

Figure 6: *(a) Illustration of candidate GCI locations distributed over a histogram based on peak amplitudes. The black pointer in (a) represents the initial threshold. The candidate GCI locations detected for different thresholds is shown in (b), (c), (d) and (e) respectively.*

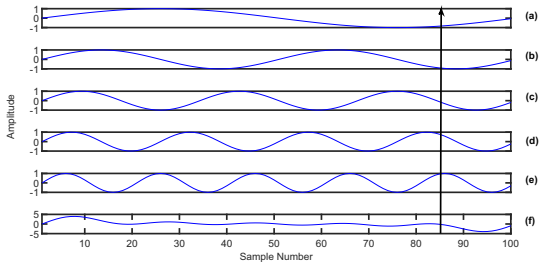

Figure 7: *Illustration of the quantitative choice of epoch interval threshold based on sum of sinusoidal harmonics. Figures (a), (b), (c), (d) and (e) represent the sinusoid of 1,2,3,4, and 5^{th} harmonics of F0 and Figure (f) represents the summation of all the harmonics of F0.*

epoch interval determined in the previous subsection in-terms of number of samples acts as the minimum distance between two epochs. In few occasions, there can be slightly higher pitched regions with epoch interval distance less than obtained threshold. An 85 % of the average epoch interval is used as minimum distance between two epochs to find the GCI locations within a voiced regions by peak picking.

A quantitative measure is validated to show that 85 % of the average epoch interval does not introduce any spurious peaks is shown in Figure 7. Filtering the signal through RCF restricts the harmonics in the signal to at most four or five harmonics, it can be observed in Figure 2(d). The ideal sinusoids of unit amplitude with five harmonics, sampling frequency of 1000 Hz and fundamental frequency of 10 Hz (100 samples) is summed to get a composite signal. The summation of five harmonics repeats after every 100 samples or 10 Hz as shown in Figure 7(f). This justifies the threshold of 85 % of the average epoch interval for reliably detecting the GCI locations in a voiced regions. Even the GCIs in lower pitched regions can be reliably detected with the obtained threshold since the distance between two epochs in general will be more than the average epoch duration, indicating lower pitch at the end of voiced segment as shown in Figure 8. From Figure 8(b), we can observer that the proposed method is succeeded in reliably detecting the GCI locations even in weakly voiced regions and irregular periodicity regions such as creaky voiced regions.

3. Experimental Results

3.1. GCI Ground Truth Creation

The GCIs can be identified as the peaks in the differenced EGG (DEGG) signal [20]. But, during vowel transitions and other

Table 1: Performance measure after adding white noise with varying SNR levels.

	Proposed Method				ZFF				SEDREAMS			
SNR(dB)	**-10**	**-5**	**0**	**5**	**-10**	**-5**	**0**	**5**	**-10**	**-5**	**0**	**5**
IDR(%)	**99.07**	**99.06**	**99.06**	**99.07**	98.27	98.26	98.26	98.25	98.97	98.93	98.91	98.94
MR(%)	0.55	0.57	0.57	0.56	**0.07**	**0.07**	**0.07**	**0.07**	0.35	0.37	0.38	0.37
FA(%)	**0.38**	**0.37**	**0.37**	**0.37**	1.66	1.67	1.67	1.68	0.68	0.7	0.7	0.69

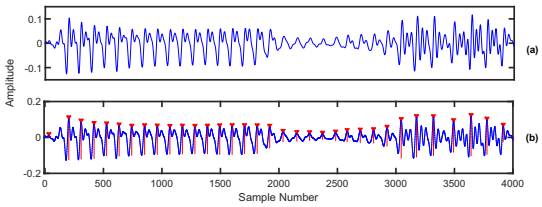

Figure 8: Illustration of the performance of the proposed GCI detection method on weakly voiced region and irregular periodicity regions. Figure (a) represents the speech signal and Figure (b) represents the detected GCIs as vertical markers.

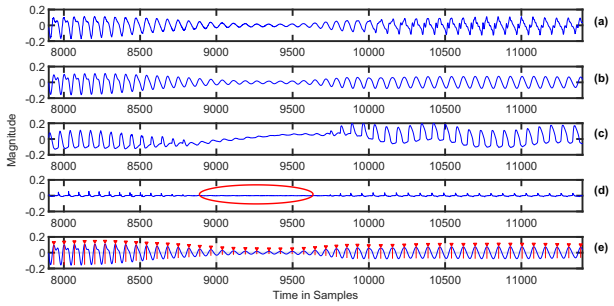

Figure 9: *Illustration of a voiced segment representing the vowel transition. Figure (a), (b), (c), (d), and (e) represent the Speech Signal, RCF Signal, EGG Signal, DEGG Signal, and GCIs detected over a RCF signal respectively. The Ellipse in (d) marks the insignificant region in DEGG Signal, even though the excitation is clear from Speech and RCF Signal.*

weakly excited voiced regions, the strength of the excitation will be weaker compared to other voiced regions resulting in a weaker peaks in the DEGG signal. Though these changes can be visually observed in the EGG signal, it is difficult to capture automatically through DEGG signal. This limitation has stimulated to create a semi-automated GCI Dataset based on EGG signal. The CMU-ARCTIC dataset [21] which consist of simultaneously recorded speech and EGG signals is used for ground truth creation and evaluation of the proposed method. The EGG signals are initially mean subtracted to remove the DC bias and first order difference is computed to obtain the DEGG signal. The DEGG is thresholded with $1/12^{th}$ of the maximum peak to obtain the initial GCI locations. The double peaks are removed and the missed GCI locations in the weakly voiced and vowel transition regions are marked manually using Sonic Visualiser, an audio analysis tool [22].

3.2. Performance Analysis

The performance of the proposed method is evaluated based on the following measures [17]: *Identification Rate (IDR)*: the percentage of glottal cycles for which exactly one GCI is detected. *Miss Rate (MR)*: the percentage of glottal cycles for which no GCI is detected. *False Alarm Rate (FA)*: the percentage of glottal cycles for which more than one GCI is detected.

Table 2: Performance measures of the proposed, ZFF and SEDREAMS methods over CMU-ARTIC Database.

Speaker	**Method**	**IDR, (%)**	**MR (%)**	**FA (%)**
BDL	Proposed	**99.16**	0.35	**0.49**
	ZFF	96.62	**0.09**	3.29
	SEDREAMS	98.44	0.41	1.15
JMK	Proposed	**99.15**	0.7	**0.15**
	ZFF	99.04	**0.09**	0.87
	SEDREAMS	98.97	0.61	0.42
SLT	Proposed	98.87	0.62	0.51
	ZFF	99.16	**0.03**	0.81
	SEDREAMS	**99.45**	0.07	**0.48**

The two prominent state-of-the-art methods for automatic detection of GCI locations based on accuracy, reliability, and robustness are ZFF and SEDREAMS. The performance measures of the proposed method are compared with ZFF and SEDREAMS as shown in Table 2. The IDR is better compared to other methods, since GCI is detected even in low excitation regions. The FA are reduced, because the proposed average pitch detection method is quite accurate. The MR is found in starting portion of the voice segment, when the pitch of the speaker is not yet stabilized. This leads to missing of GCI in the voiced segment. Further, the white noise is added over the speech signal at varying SNR levels. The results shown in Table 1 proves that the proposed method is resistant to noise.

An example of vowel transition regions where DEGG significantly fails to capture the GCI locations is shown in Figure 9. A segment of speech signal, RCF filtered, corresponding EGG, DEGG and the overlaid GCIs detected by the proposed method on the speech signal are shown in Figure 9(a), 9(b), 9(c), 9(d) and 9(e) respectively. From Figure 9(d), it can be observed that the automatic detection of GCIs from DEGG is very difficult. But, the GCIs can be easily detected based on the proposed method as shown in Figure 9(e).

4. Conclusions

A non-parametric and filtering based GCI detection method is proposed. The speech signal is filter by passing through a pulse shaping filter. The filtered signal is subjected to non-linear processing, non-parametric methods and peak picking approach to detect the GCIs. The proposed method is compared with the two state-of-the-art GCI extraction methods. The performance of the proposed method is evaluated on the ground truth GCI dataset created from CMU-ARCTIC dataset consisting of both speech and EGG signals. The experimental results showed that the proposed method is indeed immune to noise and the obtained results are better than the two state-of-the-art methods.

5. References

[1] K. Murty and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[2] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[3] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[4] B. Yegnanarayana and K. Murty, "Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[5] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1998, pp. 273–276.

[6] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[7] K. S. Rao and B. Yegnanarayana, "Prosody Modification Using Instants of Significant Excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.

[8] M. Gurunath Reddy and K. Sreenivasa Rao, "Predominant Melody Extraction from Vocal Polyphonic Music Signal by Combined Spectro-Temporal Method." in *41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(in press)*, 2016.

[9] D. Y. Wong, J. D. Markel, and A. H. Gray Jr, "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[10] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[11] T. Drugman, G. Wilfart, and T. Dutoit, "A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis." in *INTERSPEECH*, 2009, pp. 1779–1782.

[12] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979.

[13] T. Drugman and T. Dutoit, "Glottal Closure and Opening Instant Detection from Speech Signals." in *INTERSPEECH*, 2009, pp. 2891–2894.

[14] V. N. Tuan and C. d'Alessandro, "Robust Glottal Closure Detection Using the Wavelet Transform." in *EUROSPEECH*, 1999, pp. 1–4.

[15] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of Glottal Closing and Opening Instants in Voiced Speech Using the YAGA Algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[16] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[17] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of Glottal Closure Instants from Speech Signals: A Quantitative Review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[18] N. S. Alagha and P. Kabal, "Generalized Raised-Cosine Filters," *IEEE Transactions on Communications*, vol. 47, no. 7, pp. 989–997, 1999.

[19] J. G. Proakis, *Intersymbol Interference in Digital Communication Systems*. Wiley Online Library, 2001.

[20] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[21] J. Kominek and A. W. Black, "The CMU Arctic Speech Databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[22] C. Cannam, C. Landone, and M. Sandler, "Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1467–1468.