Interspeech 2018
2-6 September 2018, Hyderabad

Detection of Glottal Closure Instants in Degraded Speech using Single Frequency Filtering Analysis

G. Aneeja, Sudarsana Reddy Kadiri and B. Yegnanarayana

Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India
{aneeja.g,sudarsanareddy.kadiri}@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

Impulse-like characteristics of excitation occur at the glottal closure instant (GCI) due to sharp closure of the vibrating vocal folds in each glottal cycle. The GCIs are detected from the excitation component of the speech signal, and the excitation component is derived using inverse filtering or its variants. In this paper we propose a method for GCI detection based on single frequency filtering (SFF) of the speech signal. The SFF output has high signal-to-noise ratio (SNR) property in speech regions. The variance (across frequency) contour computed from the SFF output show rapid changes around the GCIs, and these rapid changes can be observed even when the speech signal is degraded. Thus the GCI locations can be extracted even from degraded speech using the SFF analysis. The robustness of the method is demonstrated for several cases of degradation of speech signal.

Index Terms: Glottal closure instant (GCI), single frequency filtering (SFF), zero frequency filtering (ZFF).

1. Introduction

The glottal closure instant (GCI) is the instant of significant excitation of the vocal tract, and it occurs due to rapid closure of the vocal folds in each glottal cycle. The GCI is followed by glottal closure region in a glottal cycle, which is useful to estimate the characteristics of the supraglottal vocal tract system. Knowledge of the GCI is also useful for prosody manipulation in voice conversion [1] and also in text-to-speech generation [2]. The signal around the GCI corresponds to high signal-to-noise (SNR) region within a glottal cycle, and hence features extracted around the GCIs are more robust [3]. Thus determination of GCIs from speech signals, especially when the speech is degraded is useful in several applications [4].

Several attempts have been made for extracting GCIs from speech signals [5, 4]. Among them the peaks in the error signal of linear prediction (LP) analysis have been exploited in many studies [1, 6]. Methods have been developed based on group delay functions for estimating the GCIs [7, 8, 9]. The Yet another GCI/GOI algorithm (YAGA) estimated the GCIs using the phase slope function, followed by dynamic programmimg [8]. Lines of maximum amplitude (LoMA) method uses local maximum derived from wavelet transform across multiple scales, followed by dynamic programming [10]. The multi-scale mechanism (MSM) approach relies on precise estimation of local parameters, called singularity exponent (SE) [11]. The samples with lowest SE values correspond to the GCIs. Zero frequency filtering (ZFF) is another approach proposed for GCI detection [12]. Most of these methods have also been studied for varying levels and types of degradations in the speech signals.

In this paper we propose a method of GCI detection in degraded speech based on recently proposed single frequency filtering (SFF) method [13]. The SFF output signal at each frequency will have several high SNR regions due to coherence of speech samples in a sequence and lack of coherence of noise samples. The presence of high SNR regions in the SFF outputs was exploited for speech and nonspeech detection, after suitably compensating for the noise in the degraded speech signal [13]. The SFF method was also used for extracting GCIs [14], locating burst onsets [15] and fundamental frequency extraction [16, 17]. The significance of the phase of SFF output of speech is also examined recently in [18]. In this paper the noise compensated SFF outputs are exploited for GCI detection in degraded conditions.

The paper is organized as follows. Section 2 gives an outline of the SFF method and the procedure for obtaining noise-compensated SFF outputs. Section 3 discusses the proposed method for GCI detection, which uses ZFF analysis for initial estimation. Section 4 describes the database for evaluation, and methods and evaluation for comparison. Section 5 discusses the results of evaluation. Section 6 gives a summary of the paper.

2. Single frequency filtering analysis of speech

The differenced speech signal ($x[n] = s[n] - s[n-1]$) is multiplied by the complex sinusoid $e^{j\hat{\omega}_k n}$, and the resulting frequency shifted signal $x_k[n] = x[n]e^{j\hat{\omega}_k n}$ is filtered through a single pole resonator, whose transfer function is given by [13]

$$H(z) = \frac{1}{1 + rz^{-1}}, \tag{1}$$

where $r \approx 1$, i.e., the root is close to the unit circle on the negative real axis in the z-plane. The output of the filter is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \tag{2}$$

The value of r is chosen as 0.995 in this study. Here $\hat{\omega} = \pi - \omega_k$, where $\omega_k = \frac{2\pi f_k}{f_s}$, and f_k is the desired frequency and f_s is the sampling frequency. The envelope of the k^{th} sinusoidal component of the signal is given by

$$e_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]}, \tag{3}$$

where $y_k[n] = y_{kr}[n] + jy_{ki}[n]$. The envelope $e_k[n]$ is obtained at frequencies f_k, $k = 1, 2, ..., K$, where $f_k = k\Delta f$, and Δf is the frequency spacing. For $\Delta f = 10$ Hz, the number of frequencies (K) in the interval 0-4 kHz is 400. Thus we get 400 envelopes of the SFF outputs.

As indicated in [13], the SNR of speech signal is higher in particular frequency regions and in particular time segments. For degraded speech signal, the noise power is estimated for

10.21437/Interspeech.2018-1018

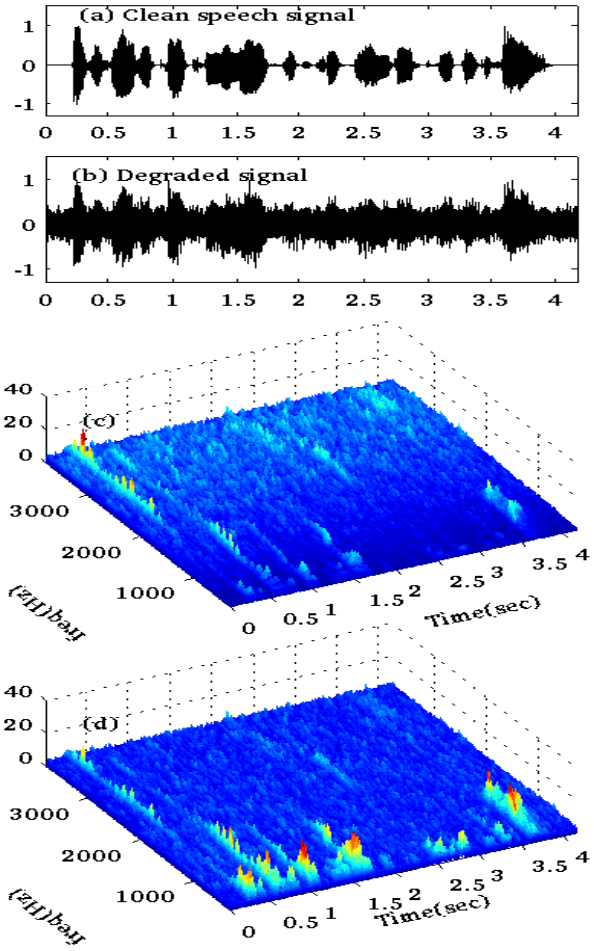

Figure 1: *(a) Clean speech signal. (b) Speech signal degraded by white noise at $SNR = 0$ dB. (c) SFF envelopes for (b). (d) Weighted SFF envelopes for (b).*

each frequency (f_k) by computing the mean (μ_k) of the lowest 20% of $e_k[n]$ values.

The normalized weight values (w_k), for noise compensation of the envelopes are given by [13]

$$w_k = \frac{\frac{1}{\mu_k}}{\sum_{l=1}^{K} \frac{1}{\mu_l}}, \tag{4}$$

where K is the number of frequencies. The noise compensated envelopes ($\hat{e}_k[n]$) are obtained by multiplying the envelopes $e_k[n]$ with the corresponding weights w_k. That is

$$\hat{e}_k[n] = w_k e_k[n], \quad k = 1, 2, ..., K. \tag{5}$$

The 20% lowest values of $e_k[n]$ are chosen under the assumption that speech utterance has at least 20% of silence. Figs. 1(c) and 1(d) show the 3-D plots of SFF envelopes and weighted SFF envelopes for clean speech signal (Fig. 1(a)) and for the signal corrupted by white noise at SNR = 0 dB (Fig. 1(b)), respectively. Note that the speech regions are emphasized in the noise compensated weighted envelopes $\hat{e}_k[n]$ (Fig. 1(d)).

3. Detection of glottal closure instants

The noise compensated envelopes $\hat{e}_k[n]$ are normalized across frequency. The normalized envelopes $\bar{e}_k[n]$ are given by

$$\bar{e}_k[n] = \frac{\hat{e}_k[n]}{\sum_{l=1}^{K} \hat{e}_l[n]}. \tag{6}$$

The variance ($\sigma^2[n]$) of the normalized envelopes is computed as follows:

$$\sigma^2[n] = \frac{1}{K} \sum_{k=1}^{K} (\bar{e}_k[n] - \mu)^2, \tag{7}$$

where $\mu = \frac{1}{K} \sum_{k=1}^{K} \bar{e}_k[n] = \frac{1}{K}$, as the envelopes are normalized across frequency.

The variance contour decreases rapidly to a minimum value around GCI [14]. Thus the slope of the variance contour is least at GCI. GCIs are detected by locating the instant of the lowest slope value of the variance contour in each glottal cycle. The slope value of the variance contour at each time instant is obtained by computing the slope of the neighboring three values. Initially, an approximate location of the GCI is obtained using the zero frequency filtering (ZFF) method [12]. If the minimum of the slope is within 2 msec of the initial estimate of GCI within a glottal cycle, then the location of the minimum slope is used as GCI, otherwise the initial estimate itself is used as GCI. This is referred as proposed method (PM) in the paper.

Figs. 2(c) and 2(d) show the variance ($\sigma^2[n]$) and slope values, respectively, computed from the clean speech envelopes $e_k[n]$. Fig. 2(a) shows the differenced electroglottograph (dEGG) signal as reference (ground truth) for locating GCIs. Notice that the variance ($\sigma^2[n]$) shows discontinuities in regions around the GCIs (Fig. 2(c)). The values of the slope have minimum values corresponding to the locations of GCIs (Fig. 2(d)).

Fig. 2(e) shows the speech signal degraded with white noise at 0 dB SNR for the corresponding clean speech in Fig. 2(b). Figs. 2(f) to 2(i) show the variance ($\sigma^2[n]$) and slope contours derived from the uncompensated envelopes ($e_k[n]$) and the noise compensated envelopes ($\hat{e}_k[n]$). Notice that the values of the variance ($\sigma^2[n]$) and slope derived from the compensated envelopes ($\hat{e}_k[n]$) show better evidence of GCIs for the degraded speech signal (Figs. 2(h) and 2(i)), when compared to the variance and slope values derived from the uncompensated envelopes ($e_k[n]$) (Figs. 2(f) and 2(g)).

The effectiveness of noise compensation for GCI detection for various types of degradations is illustrated in Fig. 3. In this figure, the slope values derived from the noise compensated envelopes ($\hat{e}_k[n]$) for five different types of degradations at 10 dB SNR are shown along with the dEGG signal as ground truth for GCI locations. Fig. 3(a) shows the dEGG signal, Figs. 3(b) to 3(f) show the slope contours derived from the variance contours for following five degradations: white, babble, machine-gun, f16, and hfchannel, respectively. From Figs. 3(b) to 3(f), it is evident that slope contours derived from noise compensated envelopes provide good evidence of locations of GCIs.

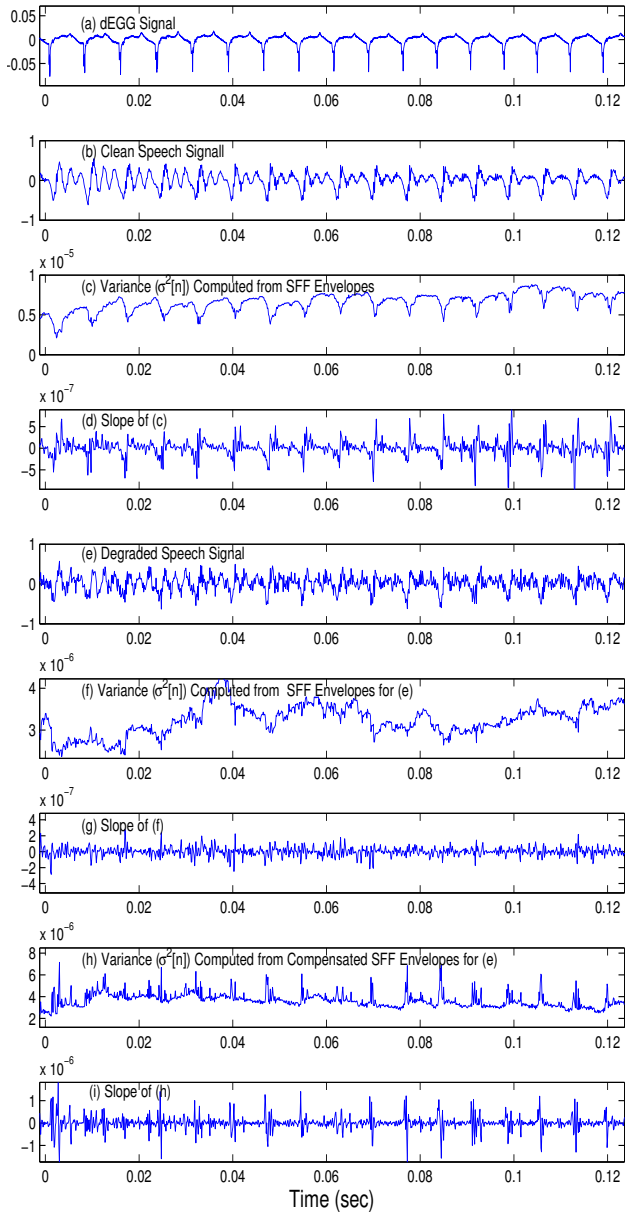

Figure 2: *(a) dEGG signal. (b) Clean speech signal. (c, d) Variance ($\sigma^2[n]$) and slope computed from SFF envelopes derived from clean speech. (e) Speech signal degraded by white noise at 0 dB SNR. (f, g) Variance and slope computed from SFF envelopes of degraded speech. (h, i) Variance and slope computed from the compensated SFF envelopes of degraded speech.*

4. Comparison of different GCI detection methods across different degradations

4.1. Database

GCI detection methods are evaluated on speech signals taken from CMU ARCTIC database [19] which contains simultaneous EGG recordings. Samples corresponding to different types of noises are taken from NOISEX database [20]. Three hundred random utterances are taken from phonetically balanced sentences, spoken by three speakers: BDL (US male), JMK (US male) and SLT (US female). The duration of each utterance is approximately 3 sec. The noises are added at SNRs of 0 dB and 10 dB. All the speech signals are downsampled to 8 kHz.

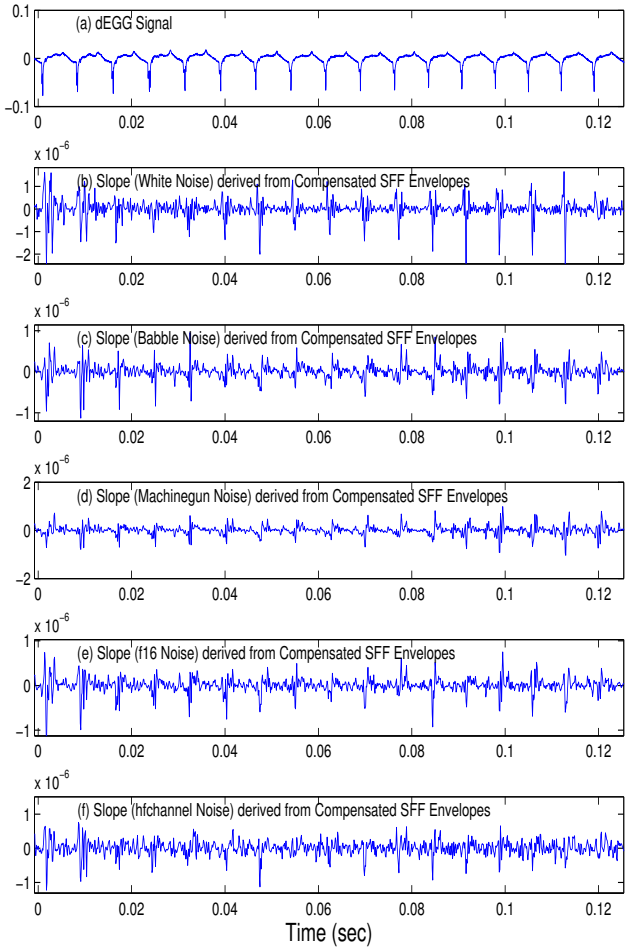

Figure 3: *(a) dEGG signal. Values of slope derived from the compensated SFF envelopes for the speech degraded at 10 dB SNR with (b) white noise, (c) babble noise, (d) machinegun noise, (e) f16 noise, and (h) hfchannel noise.*

4.2. Methods used for comparison

The following two methods are used for comparison:

- MSM method: In this, the subset of samples with lowest singularity exponent values are used to detect the GCIs. It relies on the precise estimation of multiscale parameter (singularity exponent) at each instant in the signal domain [11].
- YAGA method: In this, the information of the voice source signal (which is obtained from iterative adaptive inverse filtering (IAIF)) and stationary wavelet transform across different wavelet scales are used. The discontinuities are detected using group delay function, and the GCI candidates are measured as negative going zero crossings. The falsely detected GCIs are then removed using the M-best dynamic programming approach [8].

Table 1: *Results of GCI detection methods for different types of degradations at SNRs of 0 dB and 10 dB.*

Noise (SNR)	Method	IDR1%	MR%	FAR%	IDR2%
white(0)	PM (UW)	96.46	2.24	1.29	19.62
	PM (W)	96.35	2.32	1.33	30.81
	MSM	78.17	3.38	18.45	34.72
	YAGA	78.61	0.60	20.78	30.34
white(10)	PM (UW)	98.35	0.99	0.66	18.69
	PM (W)	98.33	1.03	0.64	47.25
	MSM	90.33	1.89	7.78	54.23
	YAGA	92.30	0.53	7.18	53.16
babble(0)	PM (UW)	89.99	2.52	7.49	19.16
	PM (W)	90.37	2.32	7.31	44.44
	MSM	81.25	3.31	15.44	35.68
	YAGA	77.39	0.87	21.74	41.57
babble(10)	PM (UW)	97.52	0.95	1.53	34.88
	PM (W)	97.71	0.90	1.39	66.27
	MSM	90.71	1.86	7.43	51.67
	YAGA	93.80	0.64	5.56	65.13
machinegun (0)	PM (UW)	92.16	5.21	2.63	51.29
	PM (W)	90.37	2.32	7.31	44.44
	MSM	81.25	3.31	15.44	35.68
	YAGA	77.39	0.87	21.74	41.57
machinegun (10)	PM (UW)	96.06	2.41	1.53	57.19
	PM (W)	96.28	2.36	1.36	78.00
	MSM	93.41	2.49	4.11	55.86
	YAGA	96.12	0.93	2.95	86.33
f16 (0)	PM (UW)	73.52	8.96	17.52	14.64
	PM (W)	74.11	8.78	17.11	36.59
	MSM	80.60	3.52	15.88	35.65
	YAGA	67.38	1.96	30.66	36.40
f16 (10)	PM (UW)	96.91	2.11	0.98	28.14
	PM (W)	97.19	2.01	0.80	60.94
	MSM	90.91	1.97	7.13	53.05
	YAGA	92.44	0.71	6.86	61.23
hfchannel (0)	PM (UW)	98.34	0.84	0.82	24.63
	PM (W)	98.33	0.85	0.81	37.00
	MSM	77.19	3.68	19.13	32.08
	YAGA	84.39	0.54	15.07	28.23
hfchannel (10)	PM (UW)	98.61	0.77	0.61	22.65
	PM (W)	98.75	0.72	0.53	52.54
	MSM	89.67	2.02	8.32	51.82
	YAGA	92.64	0.56	6.80	49.92
buccaneer1 (0)	PM (UW)	83.44	14.55	2.01	11.66
	PM (W)	83.74	14.48	1.78	25.21
	MSM	79.77	3.18	17.05	34.37
	YAGA	78.46	1.57	19.97	31.14
buccaneer1 (10)	PM (UW)	96.70	2.27	1.03	20.17
	PM (W)	96.80	2.21	0.98	52.77
	MSM	90.56	1.93	7.51	52.28
	YAGA	93.24	0.71	6.05	54.50
buccaneer2 (0)	PM (UW)	88.31	9.63	2.06	14.38
	PM (W)	88.58	9.53	1.89	28.25
	MSM	80.60	3.36	16.04	38.92
	YAGA	63.40	1.29	35.32	38.17
buccaneer2 (10)	PM (UW)	97.85	1.39	0.77	22.63
	PM (W)	97.89	1.33	0.78	54.45
	MSM	90.79	2.06	7.16	55.36
	YAGA	91.89	0.60	7.51	60.38

4.3. Evaluation measures

The following measures are used for evaluation of GCI detection methods [9].

- Identification rate1 (IDR1) : 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 (FAR): The percentage of glottal cycles for which more than one GCI is detected.
- Identification rate2 (IDR2): Identification rate1 (IDR1) within the range of -0.25 to 0.25 msec.

For better performance, IDR1 and IDR2 values should be high with low MR and FAR. The IDR2 measure indicates the percentage of correctly identified GCIs which are closer to the reference GCIs.

5. Results

The proposed method (PM) has been evaluated without and with noise compensation, indicated by PM (UW), PM (W), respectively. Table 1 shows the results obtained by the proposed methods in comparison with other methods across different types of noises at SNR levels of 0 dB and 10 dB. From the results, it can be observed that the proposed methods give comparable or better performance compared to other methods in most cases. Among the proposed methods, the noise compensation based method PM(W) has significantly increased the IDR2 value in comparison with the IDR2 value of PM(UW), while both the methods give similar IDR1 values. This is because noise compensation highlights the discontinuities due to impulse-like excitation at GCIs by reducing the spurious noise peaks. It can also be observed that the proposed methods gave good performance for different stationary and nonstationary noises in comparison with MSM and YAGA methods.

6. Summary

In this study, a method for detection of glottal closure instants (GCIs) in degraded speech conditions was proposed using single frequency filtering (SFF) method. The impulse-like discontinuities of the GCIs were exploited using the slope of the variance computed from SFF envelopes. Performance of the proposed method was compared with existing methods, and it was found that the performance is comparable or better for several cases of degradation. The performance of the proposed method improved significantly with the incorporation of noise compensation.

7. Acknowledgments

The second author would like to thank Tata Consultancy Services (TCS), India for supporting his PhD program.

8. References

[1] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[2] C. Hamon, E. Mouline, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," in *ICASSP-89*. IEEE, 1989, pp. 238–241.

[3] T. Drugman and T. Dutoit, "On the potential of glottal signatures for speaker recognition," in *Interspeech*. Citeseer, 2010, pp. 2106–2109.

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

[5] 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.

[6] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1805–1815, Dec 1989.

[7] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, Jan 2012.

[9] 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, Jan 2007.

[10] C. DAlessadro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, no. 5, pp. 601–622, 2011.

[11] V. Khanagha, K. Daoudi, and H. M. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1941–1950, Dec 2014.

[12] 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.

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

[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] B. T. Nellore, R. Prasad, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, "Locating burst onsets using sff envelope and phase information," *Proc. Interspeech 2017*, pp. 3023–3027, 2017.

[16] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *INTERSPEECH*, 2016, pp. 2155–2159.

[17] G. Aneeja and B. Yegnanarayana, "Extraction of fundamental frequency from degraded speech using temporal envelopes at high snr frequencies," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 829–838, April 2017.

[18] N. Chennupati, S. R. Kadiri, and Y. B., "Significance of phase in single frequency filtering outputs of speech signals," *Speech Communication*, vol. 97, pp. 66 – 72, 2018.

[19] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[20] 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. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/ noisex.html