Interspeech 2018
2-6 September 2018, Hyderabad

Improved Epoch Extraction from Telephonic Speech using Chebfun and Zero Frequency Filtering

B.Ganga Gowri, K.P.Soman, D.Govind

Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India

b_gangagowri@cb.amrita.edu, kp_soman@amrita.edu, d_govind@cb.amrita.edu

Abstract

Epoch in speech, represent the instant where maximum excitation at the vocal tract is obtained. Existing epoch extraction algorithms are capable of accurately extracting epoch information from clean speech signals. However, epoch extraction of band limited signals such as telephonic speech is challenging due to the attenuation of the fundamental frequency components. The present work is focused on improving the performance of epoch extraction from telephonic speech signals by exploiting the properties of Chebyshev polynomial interpolation and by reinforcing the frequency components around the fundamental frequency through the Hilbert envelope (HE). The proposed algorithm brings a refinement of the existing Zero Frequency Filtering (ZFF) method by incorporating Chebyshev interpolation. The proposed refinements to the ZFF algorithm confirmed to provide improved epoch identification rate, identification accuracy, reduced miss rate and false alarm rate. The epoch identification rate of the proposed method is observed to be better than existing methods like Dynamic Programming Phase Slope Algorithm (DYPSA), Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS), Dynamic Plosion Index (DPI) and Single Pole Filtering (SPF) methods for telephonic speech quality.

Index Terms: telephonic speech, Hilbert envelope, Chebfun system

1. Introduction

Epochs in speech signal processing represent the instants of glottal closure for voiced speech and frication or burst for unvoiced speech [1], [2]. The presence of vocal tract response makes the automatic estimation of epochs a challenging task. As the speech regions around epochs are perceptually relevant, epochal processing of speech finds important applications in speech enhancement [3], Prosody modification [4], [5], [6], speech synthesis [7] and so on. For instance, any distortion in the epoch sequence significantly affects the intelligibility of the speech signals [8]. Also, instantaneous F_0 can be used as an additional feature to improve speech recognition [9].

There are many existing methods which estimate epoch locations reliably from speech signals recorded in clean and noise free environments. Some of the popular epoch extraction methods are DYPSA, DPI, ZFF and YAGA (Yet Another GCI/GOI algorithm) [10]. Also, few other methods, process the linear prediction (LP) residual which gives an approximate representation of the glottal flow derivative for epoch extraction. SEDREAMS [11], group delay (GD) analysis and Hilbert envelope (HE) [12] based methods are such popular methods.

Even though these methods estimate epoch locations reliably, the performances of epoch extraction of these methods degrade significantly in the case of band limited signals like telephonic speech [13] or high pass filtered speech [14]. Therefore, the motive of the present work is to enhance the performance of epoch extraction from telephonic speech signals.

Telephonic channel is a band limited channel with bandwidth essentially from 300 Hz to 3.4 kHz [13], [14]. The speech signals transmitted through the telephonic channel are significantly attenuated at the low frequency components or the fundamental frequency components. Therefore, most of the methods show significant degradation in estimating the epoch locations and instantaneous F_0 contours [13]. Methods, which are devised to process the lower frequency regions in the speech are affected the most. The popular epoch extraction method ZFF, which relies on the energy of the signal around zero frequency, shows significant reduction in the performance of the epoch extraction for telephonic speech. The present work chooses the ZFF method as a case study for the improvement in the performance of epoch extraction for telephonic speech signals.

The degradation in the performance of epoch extraction for high pass filtered speech using the ZFF method is reported in [14], in which the properties of the HE in enhancing the strength of the impulse like discontinuities are exploited to improve the performance of epoch extraction using the ZFF method. The extraction of epochs from HE of speech using zero frequency filtering is also found to be effective in telephonic speech signals [15]. The epoch extraction for the telephonic speech signals can be further improved, by enhancing the strength of regions around the local pitch marks in the HE obtained from the telephonic speech by using an all pole filter prior to zero frequency filtering of the HE. The epoch identification rate for telephonic speech shows improvement for these two methods, however, they show larger deviations for the estimated epochs from the reference epochs. The present work is focused on the improvement of epoch extraction accuracies for the telephonic speech signals. Figure 1 shows the pitch contour of clean speech signal, telephonic speech signal and the HE of telephonic speech signal. The distortion in the pitch contour of telephonic speech in Figure 1(b) is due to the presence of spurious zero crossings which is reduced in the pitch contour of HE of telephonic speech signal in Figure 1(c). There are also methods like single pole filtering (SPF) proposed for improved epoch identification accuracies in telephonic speech signals.

As reported in [16], the identification accuracy of the estimated epochs is improved by processing, group delay around the peaks obtained from HE of LP residual. Motivated by this, the properties of the Chebyshev polynomial interpolation through the Chebfun system are proposed to improve the resolution of impulse like discontinuities enhanced by the HE of speech. The remaining paper is organized as follows: The description of the Chebfun system is given in Section 2, the proposed work is given in Section 3, the experiments and results are discussed in Section 4 followed by conclusion in Section 5.

10.21437/Interspeech.2018-1173

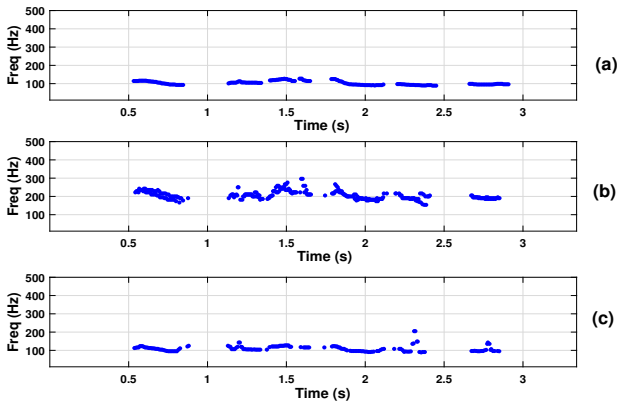

Figure 1: *Pitch contours corresponding to (a) Clean speech, (b) Telephonic speech, (c) Hilbert envelope of telephonic speech respectively*

2. Approximation of a function using Chebyshev polynomial interpolation

A function $f(x)$, defined in the interval [-1,1] can be represented by an unique polynomial interpolant $P_N(x)$ at the Chebyshev points. The Chebyshev points are defined as

$$x_j = \cos\frac{j\pi}{M}, \qquad 0 \leq j \leq M \tag{1}$$

The selection of Chebyshev points over the interval [-1,1] is illustrated in Figure 2(a). Here a semicircle passing through the end points is divided into N (here 31) arcs of equal length. Projecting the arcs on the x axis gives the Chebyshev points equivalent to (1). The Chebyshev points are unequally distributed with more points clustered at the end points when compared to the center. A Chebyshev polynomial of order k is defined as

$$T_k(x) = \cos(k \arccos(x)) \tag{2}$$

where x is the Chebyshev points. The Chebyshev polynomials of degree $k = 1, 2.., 5$ over the interval [-1,1] are shown in Figure 2(b), from which it is clear that the polynomials are fast varying at the edges than at the center. A good approximate of $f(x)$ with reduced coefficients is obtained by evaluating the Chebyshev polynomial series at the Chebyshev points rather than uniformly spaced points [17]. The Chebyshev approximation of $f(x)$ is obtained by truncating the polynomial series and is given by

$$P_N(x) = \sum_{k=0}^{N} c_k T_k(x) \tag{3}$$

where c_k is the Chebyshev coefficient. During approximation the number of coefficients required to obtain a near optimal representation of $f(x_j)$ is defined as a standard distance minimization problem [18].

$$\begin{aligned} &\min_{c} \|f - p\| \\ &s.t\, Ac = p \end{aligned} \tag{4}$$

where c is the $N+1$ Chebyshev coefficients in the range $[0, N]$, p is the polynomial values at the $M+1$ Chebyshev points and f is the function values at the Chebyshev points. The matrix A

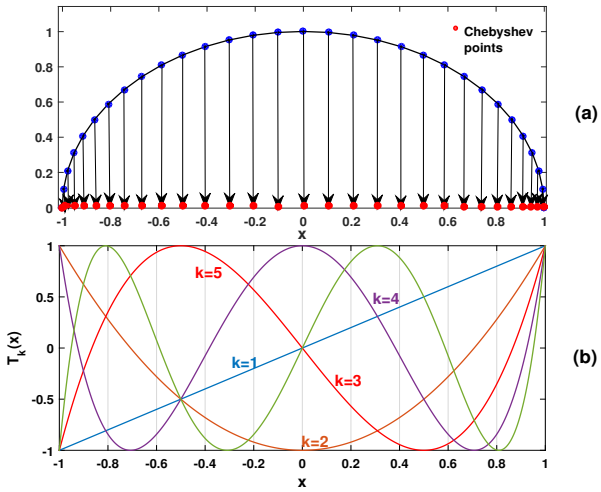

Figure 2: *(a) Selection of Chebyshev points in the interval [-1,1], (b) Chebyshev polynomials of degree k =1,2,3,4,5 (regenerated based on [19])*

consists of orthogonal Chebyshev polynomials and is given by

$$A = \begin{bmatrix} T_0(x_0) & T_1(x_0) & . & . & T_N(x_0) \\ T_0(x_1) & T_1(x_1) & . & . & T_N(x_1) \\ . & . & . & . & . \\ . & . & . & . & . \\ T_0(x_M) & T_1(x_M) & . & . & T_N(x_M) \end{bmatrix} \tag{5}$$

So, the Chebyshev approximation of $f(x)$ at the Chebyshev points is given by

$$f(x_j) \approx p_j = \sum_{k=0}^{N} c_k T_k(x_j) \quad \text{where } j = 0, 1, ..., M \tag{6}$$

The problem definition in (4) is the basis for Chebfun system which performs symbolic operations over the function values at the speed of numerical operations. Each time an operation is performed over $f(x)$, the Chebyshev coefficients are manipulated to construct the output function of the operation. In Chebfun system one can evaluate the data as if it is a continuous function. For example, operation like *sum* in MATLAB for vectors is overloaded to perform integration in Chebfun system. Similarly, zero frequency resonator operation in the Chebfun system leads to improvement in the extraction of source information at the resonator output.

3. Refined zero frequency filtering using HE and polynomial interpolation

The proposed work is an improvement of the epoch extraction using Hilbert envelope based ZFF approach [14]. The HE enhances the low frequency information in the speech signal [20]. The HE of the speech signal is approximated by the Chebfun system and is integrated twice similar to a single zero frequency resonator operation. The output signal after integration is detrended using the local mean subtraction method. The negative peaks in the de-trended signal y(n) gives the epoch estimate of the speech signal. The proposed work is described below

- The Hilbert envelope of the speech signal $s(n)$ is obtained.

$$e(n) = \sqrt{s^2(n) + s_h^2(n)} \tag{7}$$

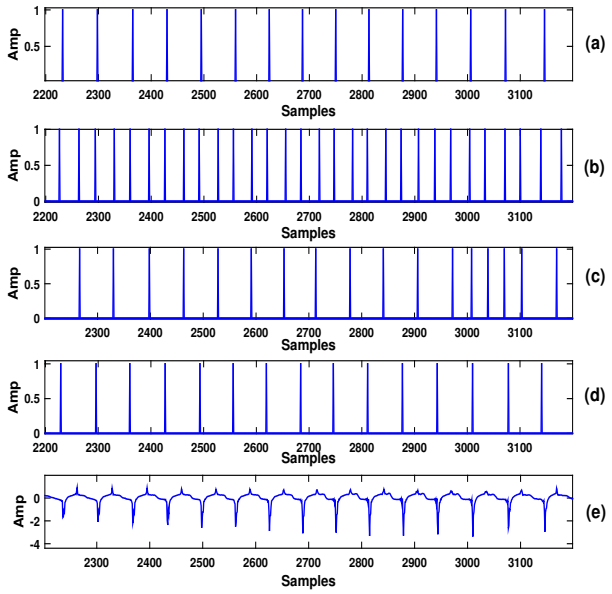

Figure 3: *Estimated epochs from (a) ZFF of Clean speech, (b) ZFF of telephonic speech, (c) ZFF of HE obtained from telephonic speech, (d) the output of proposed work obtained from the HE of the telephonic speech and (e) Differenced EGG signal*

where $s_h(n)$ is the Hilbert transform of the speech signal $s(n)$.

- The Hilbert envelope is approximated to $e(x)$ in the Chebfun system.

$$e(x) = \sum_{k=0}^{N} c_k T_k(x) \tag{8}$$

- The Chebfun approximation $e(x)$ is integrated twice similar to the zero frequency resonator operation.

$$e_1(x) = \int_a^x e(x)dx \quad ; a \leq x \leq b \tag{9}$$

$$e_2(x) = \int_a^x e_1(x)dx \quad ; a \leq x \leq b \tag{10}$$

where [a,b] is the interval over which $e(x)$ exists.

- The zero frequency resonator output, $e_2(x)$ is discretized to $e_2(n)$ and de-trended by local mean subtraction.

$$y(n) = e_2(n) - \frac{1}{2T+1} \sum_{m=-T}^{T} e_2(n+m) \tag{11}$$

where $2T + 1$ is the window length, which is fixed to achieve better performance. In general, the zero frequency resonator operation in MATLAB is carried out using the *cumsum* command twice. The same command over the data approximated in the Chebfun system performs indefinite or fractional integral operation during which the amplitudes corresponding to the higher coefficients are attenuated without reducing the overall accuracy [21], [22]. Hence, in the proposed work, an operation similar to single zero frequency resonator, instead of two is used to remove the vocal tract resonances. The negative peaks in the de-trended signal y(n) gives the epoch estimate of the speech signal. The effect of the proposed work in reducing the epoch deviation is illustrated in Figure 3. The epochs estimated using ZFF method for the clean speech signal in Figure 3(a) coincide with the reference epochs (negative peaks in first derivative of EGG signal (DEGG)) in Figure 3(e). The epochs estimated by ZFF method for the telephonic speech shows several false epochs due to the spurious zero crossings. The false epochs are reduced by using the ZFF method for the HE of telephonic speech, but the deviation from the reference has increased. This drawback has been overcome by the proposed method in which the estimated epochs shown in Figure 3(d) are closer to the reference and the estimated epochs from clean speech in Figure 3(a). Based on these visual evidences from the plots, the epoch estimation is improved using the Chebfun approximation with reduced epoch deviation.

4. Experiments and results

The proposed work is evaluated over the speech signals taken from the cmu arctic database for the speakers BDL, JMK and SLT [23]. The database contains clean speech and its corresponding EGG signals. The telephonic speech signals for evaluation is obtained by simulating the clean speech signals using the G.191 software tools given by the international telecommunication union (ITU) [24]. The clean speech signals used for epoch extraction are down sampled to 8kHz and the reference epochs for validation are obtained from the corresponding EGG signals. The performance measures used to validate the proposed work are identification rate (IDR), false alarm rate (FAR), miss rate (MR), identification accuracy (IDA) and accuracy to $\pm$0.25ms deviation. A comparison with the state of the art epoch extraction methods such as ZFF, DYSPA, DPI, SEDREAMS and SPF are made to analyze the performance of the proposed work. The SPF method gives improved epoch extraction for telephonic speech that is comparable with the clean speech. The ZFF of the HE obtained from the speech signal (HE_ZFF) is also included in the comparison, to analyze the improvement due to the Chebfun approximation. The performance measures for the clean and the telephonic speech signals by the above epoch extraction methods are given in Table 1 and

Table 1: *Performance of the epoch extraction methods for clean speech*

SPEAKER	METHOD	IDR(%)	MR(%)	FAR(%)	IDA(ms)	Accuracy to $\pm$ 0.25ms (%)
JMK	ZFF	98.5	1.19	0.33	0.48	42.9
	HE ZFF	88.36	4.81	6.83	1.6	28.1
	DYPSA	99.1	0.31	0.62	0.46	77.3
	SEDREAMS	99.6	0.3	0.1	0.5	80.1
	DPI	99.61	0.22	0.17	0.38	84.8
	SPF	94	2.8	3.4	0.88	59.5
	Proposed	98.96	0.1	0.98	0.66	67
BDL	ZFF	99.14	0.17	0.68	0.29	82.14
	HE ZFF	92.8	3.87	3.32	1.19	29.5
	DYPSA	96.27	0.84	2.88	0.43	78.4
	SEDREAMS	99.5	0.16	0.36	0.44	77.1
	DPI	99.4	0.27	0.32	0.29	89.5
	SPF	96.25	2.4	1.36	0.67	65.5
	Proposed	99.26	0.1	0.68	0.44	69
SLT	ZFF	99.8	0.11	0.04	0.31	78.9
	HE ZFF	94.9	3.02	2.05	0.69	37.8
	DYPSA	98.9	0.38	0.72	0.33	82
	SEDREAMS	99.8	0.02	0.1	0.38	76.6
	DPI	98.86	1.09	0.04	0.24	92.6
	SPF	95.87	2.82	1.31	0.59	65.4
	Proposed	98.94	0.15	0.92	0.52	68

Table 2 respectively. In case of clean speech, the IDR performance of the proposed work is compatible with the other methods used in validation and MR is better than the other methods. The performance of these measures for HE based ZFF filtering is reduced when compared to the clean speech. The same HE when given to the proposed Chebfun implementation, the measures for the latter is improved. Also the identification accuracy for $\pm$0.25ms deviation is highly improved compared to the HE_ZFF method. The performance of the epoch extraction methods such as ZFF, DYPSA, DPI and SEDREAMS are reduced for telephonic speech. The performance IDR and FAR for HE_ZFF is better compared to the ZFF method, but epoch deviation is increased. The performance measures of the proposed method for the telephonic speech show improvement when compared to the HE_ZFF method. The performance of IDR and MR values of the proposed method is comparable with that of the clean speech. The proposed work is demonstrated in Figure 4. The HE of telephonic speech signal in Figure 4(b) is closely represented by its corresponding Chebfun approximation in Figure 4(c). The proposed work output in Figure 4(d) shows a smooth signal without spurious zero crossings. The negative peaks of this signal (marked by downward arrow) gives the estimated epochs which coincide with the reference (marked by vertical dashed line) in Figure 4(e). The glottal closure timing errors of the proposed work are compared with the SPF method which gives better performance for the telephonic speech. The histogram plot of glottal closure timing errors for clean speech using SPF method and the proposed work is given in Figure 5(a), where the deviation of the proposed method is comparable with SPF method. The overall IDA at $\pm$0.25ms deviation for clean speech, obtained using SPF is 53% and using the proposed work it is 60%. The histogram plot of glottal closure timing errors for telephonic speech using SPF method and the proposed work is given in Figure 5(b), where the deviation of proposed method is little higher than the SPF method, however the performance of IDR and MR are higher for the proposed method. The overall IDA at $\pm$0.25ms deviation for telephonic speech, obtained using SPF is 67% and using the proposed work it is 69%. Overall, the proposed work shows compatible performance when compared to the SPF method for epoch extraction from clean and telephonic speech signals.

Table 2: *Performance of the epoch extraction methods for telephonic speech*

SPEAKER	METHOD	IDR(%)	MR(%)	FAR(%)	IDA(ms)	Accuracy to $\pm$ 0.25ms (%)
JMK	ZFF	10.9	2.6	86.5	1.2	5.9
	HE ZFF	86.2	8.1	5.7	1.15	4.8
	DYPSA	87.37	3.04	9.59	0.66	44
	SEDREAMS	27.8	0.8	71.4	1.2	21.1
	DPI	72.04	0.84	27.12	0.6	55.5
	SPF	95.6	2.7	1.58	0.75	50
	Proposed	98.14	0.1	1.79	0.78	47
BDL	ZFF	5.2	1.5	93.3	1.7	3.8
	HE ZFF	84	8.87	7.13	1.6	2.14
	DYPSA	87.37	3.04	9.59	0.66	34
	SEDREAMS	14.8	0.8	84.4	1.5	9.6
	DPI	87.4	1.3	11.2	0.49	63.7
	SPF	95.4	3.55	1.09	0.69	53.6
	Proposed	97.56	0.11	2.33	0.6	59.6
SLT	ZFF	66.19	6.11	27.7	1.08	30
	HE ZFF	82.7	11.9	5.4	1.7	21
	DYPSA	86.8	3.15	10.02	0.6	48.7
	SEDREAMS	91.9	1.02	7.07	0.64	53.5
	DPI	89.9	4.68	5.4	0.92	52
	SPF	91	7.62	1.7	0.67	56.3
	Proposed	96.42	0.29	3.29	0.65	65.7

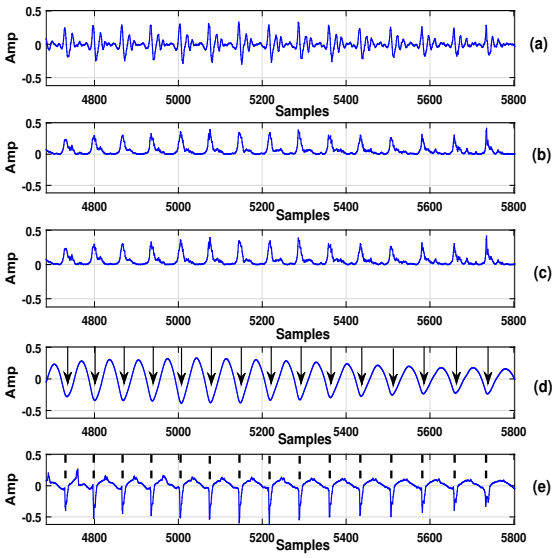

Figure 4: *The proposed epoch extraction method, (a) Speech signal, (b) HE of the speech signal, (c) Chebfun approximation of HE of speech, (d) the proposed work output obtained from the HE of the speech, (e) Differenced EGG signal and reference epoch locations (vertical dashed lines)*

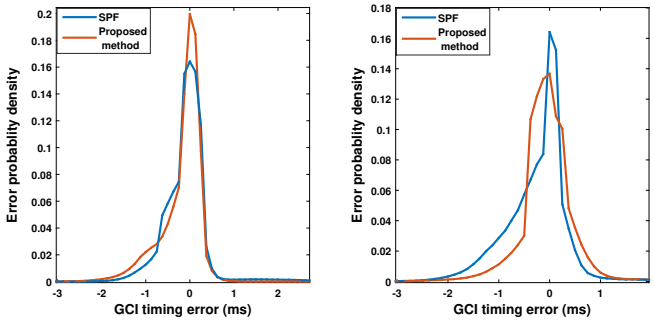

Figure 5: *Histogram of the epoch timing errors for (a) clean speech and (b) telephonic speech using SPF method (blue line) and the proposed work (orange line)*

5. Conclusion

The improvement of the epoch extraction method using the zero frequency filtering of the HE obtained from the telephonic speech signal is proposed in this work. In the HE of speech signal, the low frequency information is enhanced. The Chebyshev polynomial interpolation further enhances the discontinuities due to the epochs present in the HE of speech. The Chebfun based interpolation incorporated in the ZFF method provide better epoch identification accuracy over the HE_ZFF based epoch estimation for telephonic speech signals. The effective approximation property of the Chebyshev polynomials is exploited for efficient epoch extraction from telephonic speech signals. As a future work, the focus is on the significance of incorporating accurately estimated epoch parameters for speech recognition applications.

6. References

[1] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, 2009.

[2] 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 Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[3] P. Krishnamoorthy and S. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE trans. Audio, Speech, and Language Process.*, vol. 17, no. 2, pp. 253–266, 2009.

[4] M. R. 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 Process.*, vol. 20, no. 1, pp. 82–91, 2012.

[5] S. Prasanna, D. Govind, K. S. Rao, and B. Yegnanarayana, "Fast prosody modification using instants of significant excitation," in *Proc. Speech Prosody*, 2010.

[6] D. Govind and T. T. Joy, "Improving the flexibility of dynamic prosody modification using instants of significant excitation," *Circuits, Sys., and Signal Process.*, vol. 35, no. 7, pp. 2518–2543, 2016.

[7] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "Hmm-based speech synthesier using lf-model of the glottal source." IEEE, 2011, pp. 4704–4707.

[8] N. Adiga and S. M. Prasanna, "Significance of instants of significant excitation for source modeling." in *Proc. INTERSPEECH*, 2013, pp. 1677–1681.

[9] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2494–2498.

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

[11] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. INTERSPEECH*, 2009.

[12] L. Cohn, "Time-frequency analysis: Theory and applications," 1995.

[13] C. Vikram, S. Prasanna, C. Vikram, and S. Mahadeva Prasanna, "Epoch extraction from telephone quality speech using single pole filter," *IEEE/ACM Trans. Audio, Speech and Language Process. (TASLP)*, vol. 25, no. 3, pp. 624–636, 2017.

[14] D. Govind, S. M. Prasanna, and D. Pati, "Epoch extraction in high pass filtered speech using hilbert envelope," in *Proc. INTERSPEECH*, 2011.

[15] D. Govind, R. Vishnu, and D. Pravena, "Improved method for epoch estimation in telephonic speech signals using zero frequency filtering," in *Proc. ICSIPA*. IEEE, 2015, pp. 11–15.

[16] K. S. Rao, S. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, 2007.

[17] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*. CRC Press, 2002.

[18] M. Matus, N. Cáceres, S. Püschel-Løvengreen, and R. Moreno, "Chebyshev based continuous time power system operation approach," in *Proc. PES Genral Meeting*. IEEE, 2015, pp. 1–5.

[19] L. N. Trefethen, *Approximation theory and practice*. Siam, 2013, vol. 128.

[20] C. Langton, "Hilbert transform, analytic signal, and the complex envelope," *Signal Process. and Simulation Newsletter*, 1999.

[21] J. L. Aurentz and L. N. Trefethen, "Chopping a chebyshev series," ACM, Tech. Rep., 2015.

[22] J. Munkhammar, "Riemann-liouville fractional derivatives and the taylor-riemann series," 2004.

[23] J. Kominek and A. W. Black, "Cmu-arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[24] A. W. Black, S. King, and K. Tokuda, "The blizzard challenge 2009," in *Proc. Blizzard Challenge*, 2009, pp. 1–24.