A NOVEL FILTERING BASED APPROACH FOR EPOCH EXTRACTION

Pramod B. Bachhav, Hemant A. Patil and Tanvina B. Patel

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar-382007, India.
Email: {bachhav_pramodkumar_bhaskarrao, hemant_patil, tanvina_bpatel}@daiict.ac.in

ABSTRACT

In this paper, we propose a novel algorithm which uses simple lowpass filtering as pre-processing for detection of epochs. Lowpass filtering with an appropriate cut-off frequency removes the effect of vocal tract characteristics as formants lie in relatively higher frequency regions. The method is evaluated on entire CMU-ARCTIC database consisting of the electroglottograph (EGG) signals. Noise robustness of the proposed algorithm is evaluated in the presence of additive white noise with various SNR levels. Experimental results show that lowpass filtering make the proposed algorithm noise robust. The method gives comparable or better results with the two state-of-the-art methods, *viz.*, ZFR and SEDREAMS (which require apriori knowledge of the pitch period). In addition, the proposed method shows an improvement in *identification accuracy*.

Index Terms— Epoch extraction, lowpass filtering, glottal closure instant, group delay.

1. INTRODUCTION

An epoch is an instant of time marked by distinctive features called as *events*. In the context of speech signal, epoch is defined as the '*instant*' of significant excitation of the vocal tract system which occurs during glottal closure instant (GCI) [1]. Estimation of epochs find its applications in many areas like prosody modification, text-to-speech (TTS) synthesis, speaker recognition, emotion recognition and voice conversion. Negative peaks in the derivative of electroglottograph (EGG) are very close to the epoch locations. EGG signal is the measure of glottal airflow velocity. However, recording of EGG requires tedious lab setup. Thus, many signal processing techniques have been emerged for deriving epochs directly from pre-processed speech signal [2].

Many methods use linear predictability of speech signal as the basis for epoch estimation. Because of sudden burst of energy at GCI compared to its neighbourhood, it becomes difficult to predict speech signal around GCI. Thus, error signal obtained in Linear Prediction (LP) analysis, called as Linear Prediction Residual (LPR) supposed to contain information related to epochs. Hence, the large value of error refers to epoch location [3]. Ideally, LPR should consists of impulses near GCIs. However, there are samples of random polarity around epochs. Several studies use Hilbert envelope (HE) of the error signal for unambiguous detection of epochs [4]- [5]. An alternative to HE of LPR is proposed in [1] which is based on the global phase characteristics of minimum phase signals. The method uses positive zero-crossings of phase slope function to identify epochs. The *phase slope* function is calculated by taking average slope of the unwrapped phase of the short-time Fourier transform (STFT) of LPR as a function of time. However, this method gives rise to false alarms [6]. Therefore, the Dynamic Programming Phase Slope Algorithm (DYPSA) uses Dynamic Programming technique to select GCIs from a set of candidates to reduce false alarms [6]. Except group-delay based methods, most of the methods explained above employ *block* processing, which result in ambiguous epoch detections. In particular, the methods which rely on LPR derived by inverse filtering need selection of parameters like order of LP analysis, length of window and are dependent on energy of error signal. Zero Frequency Resonator (ZFR)-based method uses the impulsive nature of excitation for epoch extraction [7]. As discontinuity in the time-domain affects *all* frequencies, the output of ZFR should have the information of the discontinuities in the speech signal due to impulse-like excitation. The advantage of choosing zero-frequency (*0-Hz*) is that the vocal tract system has resonances around much higher frequencies than at the zero-frequency. Recently, Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) which finds rough locations of epochs from mean-based signal and then refines them through peaks of LPR, is proposed [8]. However, both ZFR and SEDREAMS need mean average pitch period of a speaker apriori. Very recently, epoch extraction using Dynamic Plosion Index (DPI) from pre-processed signal is reported in [9]. Half-wave rectified and negated integrated linear prediction residual (HWILPR) or Hilbert transform of ILPR (HTILPR) is used as the pre-processed signal. This method needs to initialize first epoch manually for every utterance. In addition, proper source (i.e., either HWILPR or HTILPR) needs to be identified for every utterance.

In this paper, we propose a novel method which passes the positive clipped and inverted speech signal through a lowpass filter for epoch extraction. Thus, the proposed method does not depend on LPR or pre-processing of LPR The details of the proposed method are discussed in next session.

2. DETAILS OF PROPOSED METHOD

2.1. Pre-processing

The voiced speech of a typical adult male have a fundamental frequency (i.e., F_o) from *85-180 Hz* and that of a typical adult female from *165-255 Hz* [10]. If speech signal is passed through a lowpass filter (with cut-off frequency (ω_n) 250 *Hz* for female voice and *180 Hz* for male voice), the filtered signal should contain pitch information. Thus, filtered signal can be approximated as sinusoidal signal with local period close to a pitch period. This observation forms basis of the proposed method. Before lowpass filtering, the speech signal was positive clipped. From Figure 1, we observe that, positive clipping just introduces DC component and does not affect frequency distribution of the signal. In particular, from Fourier series representation, we verify that the DC coefficient of a sinusoid is 0 whereas that of clipped sinusoid is $1/\pi$ and fundamental frequency of original sinusoidal is retained in the clipped sinusoid.

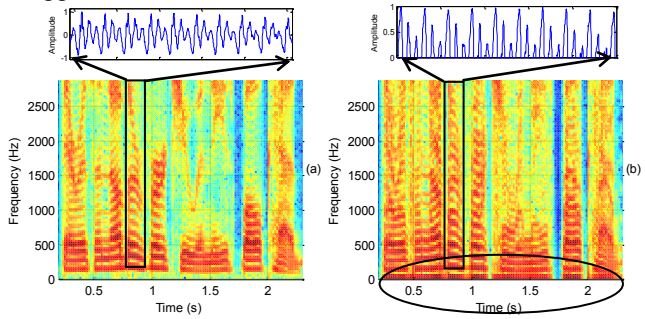

Figure 1: (a) Narrowband spectrogram of a speech signal, (b) narrowband spectrogram of positive clipped speech signal. Dotted circle indicates spectral energy density around *0-Hz*.

Thus, proposed algorithm pre-processes speech signal as follows :

- Positive portion of speech signal $x(n)$ is clipped and then it is inverted.

$$x_c(n) = \begin{cases} -x(n) & if\ x(n) < 0 \\ 0 & otherwise. \end{cases}$$

- Negated positive clipped signal $x_c(n)$ is passed through the *3rd* order Butterworth IIR lowpass filter with cut-off frequency ω_n *Hz*. $y(n)$ is the lowpass filtered signal.

Lowpass filtering introduces delay in output signal. The frequency content of $y(n)$ will be roughly in the range of 0-ω_n *Hz*. Thus, it is assumed that the *group* of frequencies present in $y(n)$ suffer from constant delay equal to the group delay of filter at ω_n. In addition, Figure 2 depicts that in frequency range 0-ω_n *Hz*, the filter almost exhibits linear phase response.

- Thus, calculate the group delay $\tau(\omega_n)$ of filter at cut-off frequency ω_n and then,
- adjust delay introduced by filtering, i.e.,

$$y(n) = y(n + \tau(\omega_n))$$

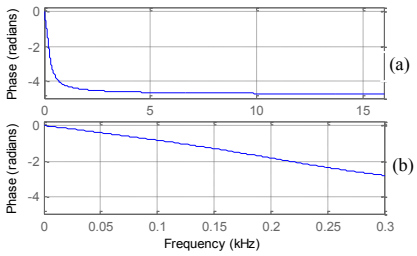

Figure 2: Phase response of filter for range (a) *0-16 kHz* ,(b) *0-300 Hz*.

Figure 3 shows a speech segment and output of lowpass filter when positive clipped and then negated speech segment passed through it. It can be observed that negative peaks in DEGG coincide with peaks of delay adjusted lowpass filtered output (solid trace). Therefore, we have experimentally observed that, after adjusting delay introduced by lowpass filter, peaks in the filtered signal correspond closely to the GCIs. After this, epochs are detected by peak picking algorithm explained in next sub-section.

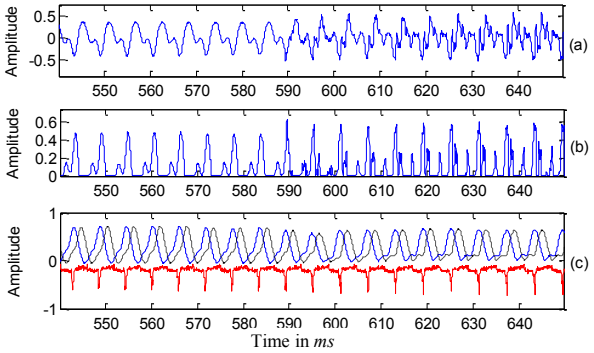

Figure 3: (a) Original speech segment, (b) positive clipped and negated speech segment (c) dashed trace : lowpass filtered output, solid trace : lowpass filtered output after adjusting delay, bottom trace: DEGG signal, negative peaks of which are taken as reference epochs.

2.2. Selection of GCIs by peak picking method

GCIs are located from peaks in filtered speech as follows :

- The peaks (E_i) in lowpass filtered speech signal with significant peak-valley difference, are selected which correspond to epoch locations.
- The peaks with small peak-valley difference were removed by applying appropriate threshold *T* (threshold *T* is selected experimentally to be *1/12th* of the average peak-valley difference in entire filtered speech signal).
- Spurious small peaks seem to appear between two major peaks in some filtered speech signals. Hence, if the peak

is *0.7* times less than either preceding (E_p) or following epoch (E_f), then it is considered spurious and is removed.

Figure 4 illustrates flowchart of the proposed algorithm for epoch estimation. Figure 5(b) shows the lowpass filtered speech consisting of few peaks with very small peak-valley difference. These peaks give rise to false alarms. Figure 5(d) shows the detected epochs with reduced false alarms.

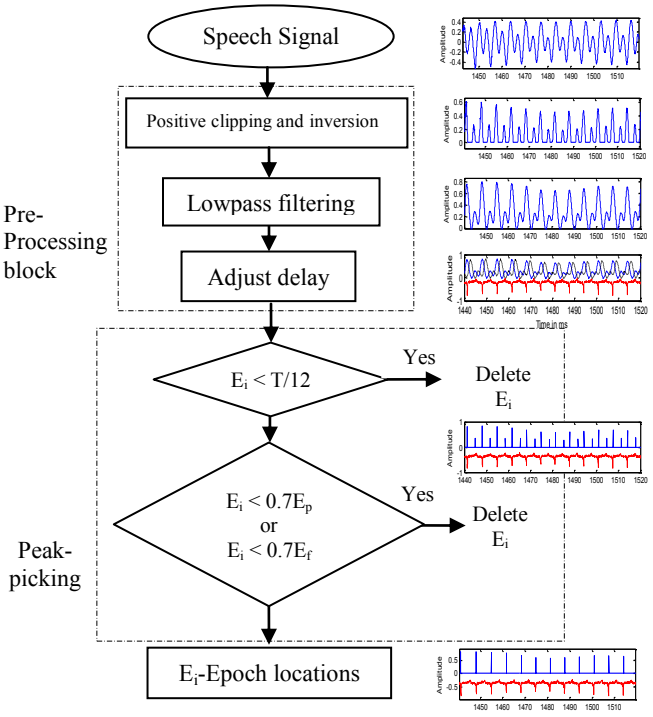

Figure 4: Illustration of the proposed algorithm for GCI estimation.

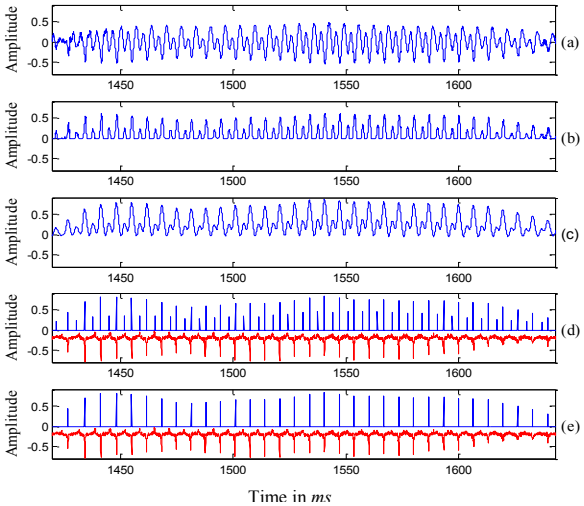

Figure 5: Selection of GCI candidates from peaks of filtered signal. (a) original speech signal, (b) positive clipped and negated speech signal (c) lowpass filtered output after adjusting delay introduced by the filter, (d) detected candidates before peak picking, (e) detected candidates after refinement of spurious peaks.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

CMU-ARCTIC database was used for evaluation of proposed method [11]. The database consists of *3377* phonetically balanced utterances of *3* speakers: SLT (US female-*1132*), JMK (Canadian male-*1114*), and BDL (US male-*1131*), digitized at *32* kHz along with EGG signals. After adjusting delay of *0.7 ms*, differenced EGG (DEGG) is used as ground truth [7]. The maximum negative peaks in DEGG are taken as reference epoch locations. The method has been evaluated on voiced segments only. Voiced-unvoiced (V-UV) decision can be made by using threshold of *1/6* times peak-to-peak value of DEGG signal [12]. Here, V-UV decision is made by applying threshold of $1/9^{th}$ of maximum negative value of DEGG signal. This is the worst case for threshold to capture low energy voiced regions [9]. Performance measures are as follows [13] :

- *Identification Rate* (IDR) : the percentage of glottal cycles for which exactly *1* 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 *1* GCI is detected;

The glottal cycles, for which exactly *1* GCI gets detected, the timing error between detected GCI and reference GCI is found.

- *Identification Accuracy* (IDA) : the standard deviation of the timing error vector. Small value of IDA corresponds to higher identification rate.
- *Accuracy to ± 0.25 ms* : The percentage of detections for which the timing error is less than ± *0.25 ms*.

3.2. Performance on clean speech

Table 1 shows the performance of ZFR, SEDREAMS and the proposed method on CMU-ARCTIC database.

Table 1: Comparison of results over CMU-ARCTIC database

Speaker	Method	IDR (%)	MR (%)	FA (%)	IDA (ms)	Acc. to ± 0.25 ms(%)
BDL	Proposed*	98.43	0.98	**0.59**	**0.34**	63.27
	Proposed-WC	91.89	0.28	7.83	0.35	60.83
	Proposed-NC	84.44	0.76	14.8	0.75	34.96
	ZFR	96.62	**0.09**	3.29	0.38	74.76
	SEDREAMS	**98.44**	0.41	1.15	0.42	**81.81**
JMK	Proposed*	97.54	2.18	**0.28**	**0.56**	47.62
	Proposed-WC	98.14	1.34	0.52	0.59	45.06
	Proposed-NC	96.70	1.26	2.04	0.56	36.24
	ZFR	**99.04**	**0.09**	0.87	0.66	34.72
	SEDREAMS	98.97	0.61	0.42	0.65	**65.59**
SLT	Proposed*	98.93	0.44	0.64	**0.28**	69.12
	Proposed-WC	96.66	0.19	3.15	0.38	60.51
	Proposed-NC	95.77	0.67	3.56	0.66	25.56
	ZFR	99.16	**0.03**	0.81	0.35	**78.67**
	SEDREAMS	**99.45**	0.07	**0.48**	0.32	72.99

* Proposed method with positive clipping.

Table 2: Comparison of epoch extraction techniques for additive white noise on CMU-ARCTIC database at various SNR levels.

	Proposed				Proposed-WC				ZFR				SEDREAMS			
SNR(dB)	-10	-5	0	5	-10	-5	0	5	-10	-5	0	5	-10	-5	0	5
IDR (%)	98.29	98.30	98.30	98.30	95.50	95.50	95.50	95.52	98.27	98.26	98.26	98.25	**98.97**	**98.93**	**98.91**	**98.94**
MR (%)	1.17	1.17	1.18.	1.19	0.59	0.59	0.60	0.60	**0.07**	**0.07**	**0.07**	**0.07**	0.35	0.37	0.38	0.37
FA (%)	**0.54**	**0.53**	**0.52**	**0.51**	3.90	3.91	3.90	3.88	1.66	1.67	1.67	1.68	0.68	0.70	0.70	0.69
IDA (%)	**0.39**	**0.39**	**0.39**	**0.39**	0.44	0.44	0.44	0.44	0.48	0.48	0.49	0.5	0.50	0.49	0.48	0.47
Acc.to ±*0.25* ms (%)	59.98	59.98	60	60	55.46	55.46	55.47	55.47	62.17	62.21	62.20	62.14	66.36	69.15	71.16	72.78

We compare our results with ZFR and SEDREAMS as both methods perform better than HE-based, group delay-based methods and DYPSA [7], [8]. We investigate importance of positive clipping by evaluating performance of the proposed method without positive clipping (Proposed-WC) and with negative clipping (Proposed-NC). IDR and FA for the proposed method are almost comparable with that of ZFR and SEDREAMS. The selection of proper threshold may help to improve MR. The proposed method gives better IDA over these two recently proposed methods. The advantage of proposed method over other two is its simplicity. It does not require prior mean pitch period information as required in existing ZFR and SEDREAMS methods.

3.3. Performance on signal degradation conditions

The method has been tested and compared on noisy speech signals. For this purpose, white noise was added to the speech signals from CMU-ARCTIC database at different Signal-to-Noise Ratio (SNR) levels. The performance measures were averaged over all *3* speakers available in the database. The SNR is varied from *-10* dB to *5* dB in steps of *5* dB. The white noise was taken from NOISEX-92 [14]. database. Table 2 displays comparison of the proposed method over ZFR and SEDREAMS at various SNR levels. It is observed that the proposed method performs well in terms of FA and IDA and gives comparable results in terms of IDR over other two methods under severe degraded conditions.

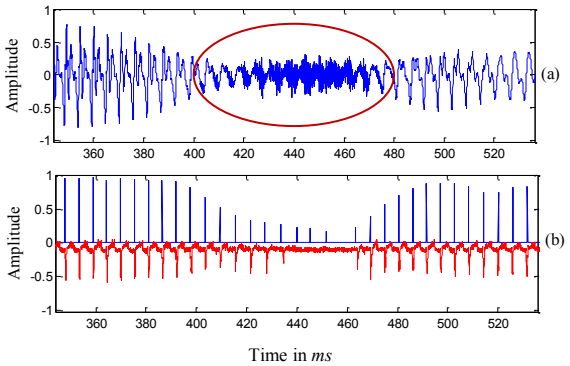

Figure 6: Illustration of proposed method for voiced fricative. (a) a speech segment with a voiced fricative /z/, (b) top trace- estimated epoch locations, bottom trace- DEGG signal.

3.4. Analysis of proposed method in voiced fricatives and low voicing regions

The proposed method detects epochs in low voicing regions as well as voiced fricatives. In Figure 6, a segment of speech with a vowel followed by a voiced fricative followed by a vowel, is shown. It is observed that, epochs in the voiced fricative are detected properly. Thus, it is clear that algorithm detects epochs in all voiced sounds. Figure 7 depicts the performance in low voicing regions. It means that the proposed method does not depend on energy of vowel and epochs in low voicing regions can be detected.

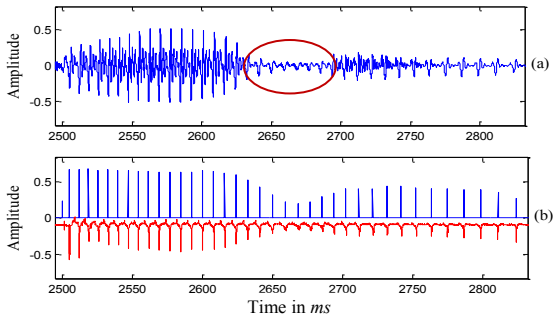

Figure 7: Illustration of proposed method for low voicing region. (a) a speech segment with low voicing region, (b) top trace-estimated epoch locations, bottom trace- DEGG signal.

4. SUMMARY AND CONCLUSIONS

A simple method for epoch extraction has been proposed in this paper. Performance of the method has been reported comparable or better with other existing techniques. Future work will be emphasized on reducing MR by selection of proper threshold. Reduction in MR ultimately will help to improve other performance measures. Simple lowpass filtering makes the proposed method immune to severe degraded conditions. The peak-valley difference of lowpass filtered output may be extended to characterize strength of excitation (SOE) in a glottal cycle which authors would like to investigate in future.

5. ACKNOWLEDGEMENTS

The authors would like to thank Department of Electronics and Information Technology (DeitY), Govt. of India for sponsoring the consortium project. They also thank the authorities at DA-IICT, Gandhinagar, India.

6. REFERENCES

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

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

[3] Bishnu S. Atal and Suzanne L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 28, pp. 637-655, 1971.

[4] K. Sreenivasa Rao, S. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 762-765, 2007.

[5] T. 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.

[6] Patrick A. Naylor, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 34--43, 2007.

[7] K. Sri Rama Murty and B. Yegnanarayana, "Epoch Extraction From Speech Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 1602-1613, November 2008.

[8] Thomas Drugman and Thierry Dutoit, "Glottal closure and opening instant detection from speech signals.," in *INTERSPEECH*, 2009, pp. 2891-2894.

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

[10] Ingo R. Titze and Daniel W. Martin, "Principles of voice production," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1148-1148, 1998.

[11] "CMU-ARCTIC Speech Synthesis Databases," [Online]: http://festvox.org/cmu_arctic/index.html (Last Accessed: September 25, 2014).

[12] Donald G. and Ahn, Chieteuk Childers, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505-519, 1995.

[13] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry 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.

[14] "Noisex-92,"[Online]: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html (Last Accessed: September 25, 2014).