Epoch Extraction in High Pass Filtered Speech using Hilbert Envelope

D. Govind , S. R. M. Prasanna and Debadatta Pati

Department of Electronics and Electrical Engineering, IIT Guwahati, Assam, India

{dgovind, prasanna, debadatta}@iitg.ernet.in

Abstract

Hilbert envelope (HE) is defined as the magnitude of the analytic signal. This work proposes HE based zero frequency filtering (ZFF) approach for the extraction of epochs in high pass filtered speech. Epochs in speech correspond to instants of significant excitation like glottal closure instants. The ZFF method for epoch extraction is based on the signal energy around the impulse at zero frequency which seems to be significantly attenuated in case of high pass filtered speech. The low frequency nature of HE reinforces the signal energy around the impulse at zero frequency. This work therefore processes the HE of high pass filtered speech or its residual by zero frequency filtering for epoch extraction. The proposed approach shows significant improvement in performance for the high pass filtered speech compared to the conventional ZFF of speech.

Index Terms: Epochs, zero frequency filtering, Hilbert envelope

10.21437/Interspeech.2011-520

1. Introduction

Epochs in speech correspond to the instants of significant excitation of the vocal tract [1][2]. The instants of significant excitation are impulse in nature having significant amplitude in a local sense. They represent events like glottal closure and glottal opening in case of voiced speech and onset of burst and frication in case of unvoiced speech. The present work focusses on the extraction of glottal closure instants. The knowledge of epoch locations is very useful in many speech applications like extraction of vocal tract system parameters, prosody modification and speech enhancement [3, 4, 5]. Due to its significance, many methods have been proposed in the literature for epoch extraction [6, 1, 7, 2]. Among these methods, zero frequency filtering (ZFF) method provides comparatively better epoch estimation performance[2].

The property of impulse like discontinuity is exploited in ZFF method. The time domain representation of impulse function has an equivalent frequency domain representation of impulses uniformly located at all the frequencies including zero frequency, separated by fundamental frequency, forms the basis for ZFF method [2]. In ZFF method, speech is passed through a resonator located at the zero frequency which preserves the signal energy around the impulse present at zero frequency and removes all other information, mainly due to the vocal tract resonances. The trend in the output of the zero frequency resonator is removed further by considering a window of length one to two pitch periods and the trend removed signal is termed as the zero frequency filtered signal [2]. The positive zero crossings of the zero frequency filtered signal give the location of epochs. This method works well as long as the speech signal has significant energy around the impulse at zero frequency. However, in case of high pass or band pass filtered speech like telephone bandwidth, the signal energy around the impulse at zero frequency is significantly attenuated. Under such conditions, ZFF of speech may tend to give unreliable or spurious epoch locations.

The signal energy of the impulse around the zero frequency can be reinforced by using the Hilbert envelope (HE). The HE is defined as the magnitude of the analytic signal or complex time function [6, 8]. The basic property of the analytic signal is its single sided positive spectrum. To ensure this, the HE accommodates relatively low frequency components and the phase part of the analytic signal accommodates relatively high frequency components [8]. The low frequency components of the HE in turn reinforces the energy around the impulse at zero frequency. As a result, the ZFF of HE should identify the epochs with better accuracy and hence the motivation. This work further shows that the reinforcement of signal energy around the zero frequency is better in case of speech than its linear prediction (LP) residual. The organization of the paper is as follows: Section 2 describes the ZFF of speech for estimating the epochs and demonstrates its limitation in case of high pass filtered speech. Section 3 proposes the ZFF of HE for high pass filtered speech. Section 4 describes the evaluation of the proposed method on telephone speech. Finally, Section 5 summarizes the work with a mention on the possible future work.

2. Zero Frequency Filtering of high pass filtered speech

The algorithmic steps to estimate the epochs in speech by ZFF are as follows [2]:

- Difference input speech signal $z(n)$

$$x(n) = z(n) - z(n-1) \tag{1}$$

- Compute the output of cascade of two ideal digital resonators at 0 Hz

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{2}$$

  where a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1

- Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{3}$$

  where $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{n=-N}^{N} y(n)$ and $2N+1$ corresponds to the average pitch period computed over a longer segment of speech

- The trend removed signal $\hat{y}(n)$ is termed as zero frequency filtered signal.

- The positive zero crossings of the filtered signal will give the location of the epochs.

Copyright © 2011 ISCA

Figure 1(a) plots a segment of clean speech. The corresponding output from ZFF and detected epochs are shown in Figures 1(c) and (e), respectively. The output ZFF shows proper zero crossings and hence the detected epochs are accurate. In the next case let us consider an high pass filtered speech obtained by passing the clean speech through an high pass filter with an arbitrary cut-off frequency of 500 Hz. The choice of cut off frequency of 500 Hz is made so as to exclude the normal human (male and female) pitch range. The corresponding plots of the above process are shown Figures 1(b), (d) and (f), respectively. As it can be observed, the output of ZFF shows spurious zero crossings at some places and hence the detected epochs. Since the only change is high pass filtering of speech, this may be attributed to the ZFF approach followed for detecting the epochs. Thus we may say that, the poor performance in case of high pass filtered speech is due to the significant attenuation of signal energy around the impulse at zero frequency.

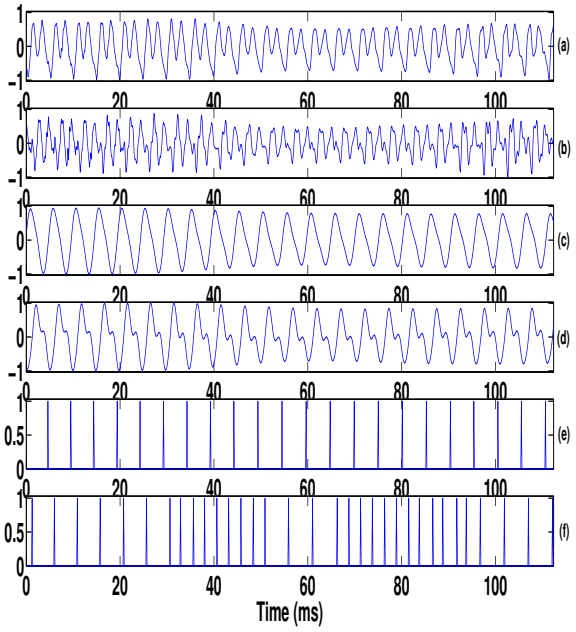


Figure 1: Epoch detection by ZFF. (a) clean speech, (b) high pass filtered speech. Output of ZFF from (c) clean speech and (d) high pass filtered speech. Estimated epoch locations from (e) clean speech and (f) high pass filtered speech.

The extent of degradation in the performance can be obtained by comparing the detected epochs with the reference epoch locations indicated by the impulse like discontinuities in the differentiated Electroglottogram (EGG) of the same speech segment as performed in [2]. For evaluating the estimated epochs from the speech, following measures are used [7].

- Larynx cycle: The range of sample (1/2) $(l_{r-1} + l_r) < n < (1/2)(l_{r+1} + l_r)$ where l_r, l_{r-1} and l_{r+1} are the current, preceding and succeeding reference epoch locations respectively
- Identification Rate (IDR): The percentage of larynx cycles for which exactly one instant of significant excitation is detected.
- Miss Rate (MR): The percentage of larynx cycles for which no epoch is detected.

Table 1: Performance evaluation of the epoch estimation using ZFF and DYPSA of clean and high pass filtered speech using 500616 reference epochs. HPF refers to high pass filtered.

Signal	IDR (%)	MR (%)	FAR (%)	IDA (ms)
ZFF				
Speech	99.54	00.02	00.44	0.2923
HPF Speech	92.43	00.05	07.52	0.2825
LP residual of HPF speech	96.95	00.09	02.96	0.3116
HE of HPF Speech	98.96	00.49	00.56	0.4913
HE of LP residual of HPF Speech	96.21	01.47	02.32	0.5973
DYPSA				
Speech	98.05	0.62	1.34	0.3481
HPF Speech	92.72	1.90	5.38	0.3717

- False Alarm Rate (FAR): The percentage of larynx cycles for which more than one epoch is detected.
- Identification Error (ζ): The timing error between the reference and detected instants of significant excitation in larynx cycles for which exactly one instant of significant excitation was detected.
- Identification Accuracy (σ) (IDA): The standard deviation of the identification error ζ. Small values of σ indicate high accuracy of identification.

The performance is evaluated over a total of 500616 reference epochs obtained from EGG recordings of 3 speakers (two males and one female) having 1132 phonetically balanced sentences from CMU Arctic database [9]. Epoch estimation performance for various measures mentioned above are given in the Table 1. The higher epoch identification rate and reduced miss rate, false alarm rate and identification errors when compared to the reference epochs indicate the superiority of the ZFF method in accurately estimating the epochs from clean speech. The epochs estimation performance of DYPSA, which is another popular method for epoch estimation, is also provided in Table 1. The improved performance of ZFF compared to DYPSA is reported in [2]. All these speech signals are passed through an high pass filter with cut-off frequency of 500 Hz and then used for epoch detection. The performance for high pass filtered (HPF) speech is also shown in Table 1. The miss rate seems to be same, but the FAR increases significantly resulting in lower IDR. The same trend can be observed in case of epoch estimated using DYPSA method also. It has to be noted that there is an improvement in the epoch estimation performance for the LP residual of high pass filtered speech as compared to the ZFF of high pass filtered speech. This is because of the sharp discontinuities of the residual samples at epoch locations as compared to the speech signal. As the LP residual is derived from the high pass filtered speech, it still preserves the HPF characteristics. That is, there is no reinforcement of energy at the zero frequency. As a result the epoch estimation performance is degraded as compared to that of clean speech.

3. Hilbert envelope based epoch extraction from high pass filtered speech

In order to improve the epoch detection performance, the high pass filtered speech has to be transformed such that the signal energy around the impulse at zero frequency is enhanced. For this we propose the use of Hilbert envelope (HE) [6, 8].

3.1. Hilbert Envelope

Let $s_a(n)$ be the analytic signal of a given signal $s(n)$. Then,

$$s_a(n) = s(n) + js_h(n) \tag{4}$$

where $s_h(n)$ is the Hilbert transform of $s(n)$. The Hilbert transform is computed as

$$s_h(n) = IDTFT(S_H(\omega)), \tag{5}$$

where

$$S_H(\omega) = \begin{cases} +jS(\omega), -\pi \leq \omega < 0 \\ -jS(\omega), 0 \leq \omega \leq \pi \end{cases} \tag{6}$$

and $S(\omega)$ is the DTFT of $s(n)$. DTFT refers to discrete time Fourier transform and IDTFT refers to inverse of DTFT.

Let $h(n)$ be the HE. It is defined as the magnitude of $s_a(n)$ i.e.,

$$h[n] = |s_a(n)|\,. \tag{7}$$

Therefore,

$$h(n) = \sqrt{s^2(n) + s_h^2(n)}. \tag{8}$$

Let $\phi(n)$ be the phase of $s_a(n)$. It is defined as

$$\phi(n) = \tan^{-1}\left(\frac{s_h(n)}{s(n)}\right). \tag{9}$$

Figure 2 shows $s(n)$ taken as a speech segment, its $h(n)$ and $\phi(n)$. $h(n)$ is unipolar in nature, since it is a magnitude function and $\phi(n)$ is bipolar in nature.

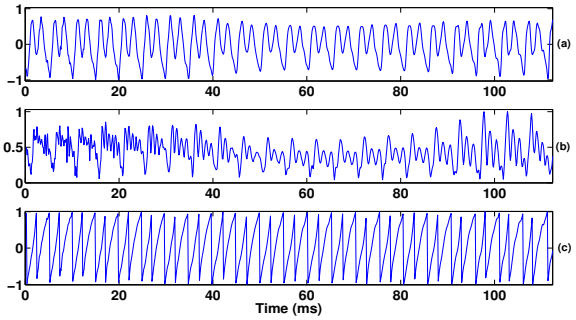


Figure 2: (a) Clean speech, (b) HE and (c) phase of its analytic signal.

The fundamental characteristic of $s_a(n)$ is, its spectrum $S_A(\omega)$ is one sided, i.e., exists only in the range $0 \leq \omega \leq \pi$. To ensure this always, whenever $s_a(n)$ is computed for given signal $s(n)$, only relatively low frequency components are placed in $h(n)$ and high frequency components are placed in $\phi(n)$ [8]. Otherwise, it will violate the single sided spectrum property of $s_a(n)$ and it ceases to be complex time function. For instance, if a given signal $s(n) = cos(\omega_1 n) + cos(\omega_2 n)$ and $\omega_1 < \omega_2$, then $s_a(n) = cos(\omega_1 n)e^{j\omega_2 n}$.

3.2. Epoch extraction from HE of HPF Speech

In case of high pass filtered speech, the relatively low frequency nature of HE enhances the signal energy around the impulses at all frequencies, including the impulse at zero frequency. This property is exploited for the detection of epochs. Figure 3(a) shows a segment of high pass filtered speech (same as Figure 1(b)) and its HE in Figure 3(b). The presence of strong impulses in case of HE may be observed by comparing these two figures which may be attributed to the nature of HE. Figure 3(d) shows the output of ZFF of the HE and the detected epochs in Figure 3(e). These epochs may be compared with those from speech given in Figure 3(f) (same as Figure 1(f)). As it can be observed, the spurious peaks are eliminated in the case of HE.

The performance is evaluated over a total of 500616 reference epochs in the Arctic database for the high pass filtered case using the HE of HPF speech and the various values are given in Table 1. The HE of HPF speech provides significantly better performance compared to the HPF speech demonstrating its robustness to high pass filtering.

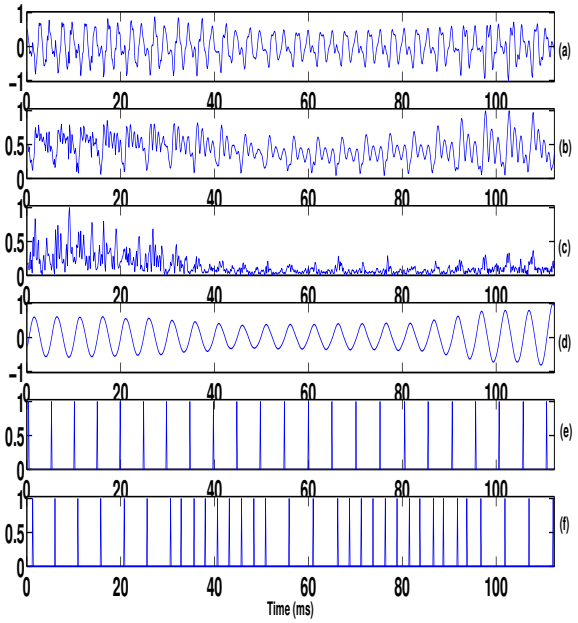


Figure 3: (a) HPF speech (b) HE of HPF speech and (c) HE of LP residual of HPF speech. Output of ZFF from (d) HE of HPF speech. Estimated epoch locations from ZFF of (e) HE of HPF speech and (f) HPF speech.

3.3. Epoch Extraction from Hilbert Envelope of LP Residual of HPF Speech

The impulse like discontinuities can also be located from the Linear Prediction (LP) residual which is obtained by the LP analysis of speech [10]. The speech is processed by the 10^{th} order LP analysis to obtain the LP residual and then HE of the LP residual is computed. The epochs are determined by the ZFF of the HE of the LP residual from the HPF speech. Table 1 shows improvement in the performance of the estimated epochs using ZFF of HE of LP residual over the high pass filtered speech, but the performance is lower compared to the HE

Table 2: Performance evaluation of the epoch estimation using ZFF method for NTIMIT telephonic database.

Signal	IDR (%)	MR (%)	FAR (%)	IDA (ms)
Speech	51.75	04.55	43.70	0.8477
Residual	53.10	04.45	42.50	0.8341
HE of Speech	79.74	16.17	04.10	0.8200
HE of LP residual	76.60	16.16	07.24	0.8246

of speech. The reason for the improved performance is due to the low frequency nature of the HE of the LP residual. The better performance of HE of Speech may be attributed to the better reinforcement of the epochal information while computing the analytic signal. This can be observed by comparing Figures 3(b) and (c). The epochal information seem to be much stronger in case of HE of speech compared to the other.

4. Hilbert envelope based epoch extraction from telephone speech

The main application of epoch extraction in HPF speech can be found in telephone speech. NTIMIT telephonic speech database is used for the study. For evaluating the performance, the epochs from the corresponding speech signals in the TIMIT database are taken as reference. The evaluation is done over the whole NTIMIT databse training set having 462 speakers. The reference epochs are derived using the ZFF filtering of corresponding speech files from the TIMIT database training set and in total we have 683414 reference epochs.

The Table 2 shows that there is significant improvement in the epoch estimation performance using ZFF of HE of telephonic speech. The performance of the epoch parameters obtained by the ZFF of LP residual HE is also significant compared to the ZFF of telephonic speech. Both these performances demonstrate the robustness of HE for epoch extraction. Figure 4 plots the pitch computed using ZFF of TIMIT speech, NTIMIT speech, HE of NTIMIT speech and HE of LP residual of NTIMIT speech, respectively. As shown in figure, the pitch values obtained from the HE of NTIMIT speech represents smooth contour and closely resembles that of the TIMIT speech. The increased false epoch detection is responsible for the scattered pitch values in the case of ZFF of NTIMIT speech. The pitch contour computed from the HE of LP residual of NTIMIT speech is also smooth, but has relatively more spurious values compared to that of HE of NTIMIT speech.

5. Summary and future work

This work demonstrated the limitation of ZFF of speech for epoch extraction from high pass filtered speech. The work later proposed a method based on ZFF of HE of speech for epoch extraction in case of HPF speech. The proposed approach seem to be superior even compared to the ZFF of HE of LP residual. Finally, the significance of proposed approach is demonstrated in case of NTIMIT telephone speech. In all the cases, ZFF of HE of speech provided the best performance and therefore we propose the use of HE of speech for epoch extraction in case of high pass filtered speech.

The superior performance of HE shows its robustness compared to speech. However the temporal resolution in case of HE is relatively poor compared to speech for HPF case. The future work should focus on improving the temporal resolution.

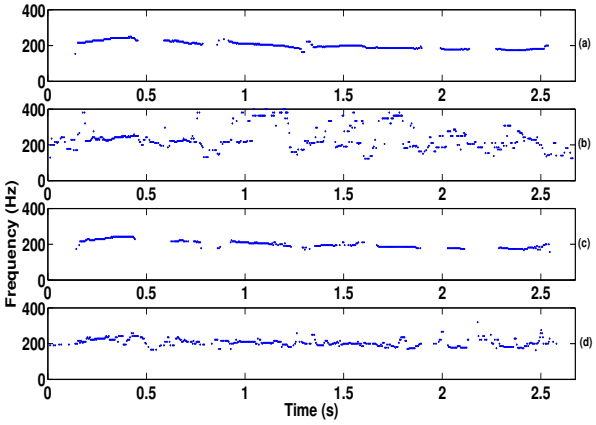


Figure 4: Pitch contours derived from (a) TIMIT speech, (b) NTIMIT speech, (c) NTIMIT speech HE and (d) NTIMIT LP residual HE.

6. Acknowledgements

This is part of ongoing (2007-2011) UKIERI project between IIT Guwahati, India, IIIT Hyderabad, India and CSTR, University of Edinburgh, UK.

7. References

[1] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 4, pp. 325–333, Sep.1995.

[2] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1614, Nov. 2008.

[3] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 4, pp. 313–327, July 1998.

[4] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 972–980, May 2006.

[5] E. A. P. Habets, N. D. Gaubitch, and P. A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *in Proc. ICASSP*, Jan. 2008, pp. 4577–4580.

[6] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, 1979.

[7] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using DYPSA algorithm," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[8] L. Cohen, *Time-Frequency Analysis: Theory and Applications*, S. P. Series, Ed. ser. Signal Processing Series. Englewood Cliffs: Prentice-Hall, 1995.

[9] J. Kominek and A. Black, "CMU-Arctic speech databases," in *in 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224.

[10] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.