Significance of Instants of Significant Excitation for Source Modeling

Nagaraj Adiga and S. R. Mahadeva Prasanna

Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati,
Guwahati 781039, India.
{ nagaraj, prasanna } @iitg.ernet.in

Abstract

The objective of this work is to demonstrate the significance of instants of significant excitation for source modeling. Instants of significant excitation correspond to the glottal closure, glottal opening, onset of burst, frication and a small number of excitation instants around them. The speech signal is processed independently by zero frequency filtering (ZFF) to obtain epochs. The epochs are used as anchor points for extracting the instants of significant excitation from different representations of speech. The different representations include sequence of strength weighted epochs, small range of samples around epochs from the linear prediction (LP) residual, Hilbert envelope (HE) of LP residual and the cosine of phase sequence. The strength weighted epoch sequence generates speech which is intelligible, but synthetic in nature. By considering a small region of instants of significant excitation around the epochs, the naturalness of synthesized speech increases significantly.

Index Terms: epochs, instants of significant excitation, LP residual, Hilbert envelope and cosine of phase.

10.21437/Interspeech.2013-29

1. Introduction

The source modeling is important in both speech synthesis and coding [1, 2]. The intelligibility, naturalness and prosodic information are well preserved in the better source modeled case. The speech synthesized using simple two state source model is intelligible, but metallic and suffers from absence of naturalness and prosodic information [1]. Alternatively, use of complete linear prediction (LP) residual as the source signal provides speech signal nearly equivalent to original speech [3], but needs more data rate and also less amenable to modeling. Hence a via media is to be worked out for better source modeling at reduced data rate to preserve naturalness and prosodic information. There are many attempts in the speech coding area for achieving the same [3]. All these studies focus on better modeling of information present in the LP residual at reduced data rate [1]. The present work aims to demonstrate the significance of instants of significant excitation for source modeling [4].

In the existing literature [4–6] , instants of significant excitation or more commonly termed as epochs, correspond to glottal closure instants (GCIs), glottal opening instants (GOIs), onset of burst and frication [4]. *The present work extends the definition of instants of significant excitation to include a small number of additional excitation instants around them also. Thus instants of significant excitation refers to epochs plus a small range of excitation instants present around them.* The present work proposes to detect the epochs from speech by any epoch extraction method and use them as anchor points for selecting the instants of significant excitation and demonstrates their significance for source modeling. In particular, the interest is to know the type of representation and also amount of excitation instants to be considered around the epochs as instants of significant excitation to preserve naturalness and prosodic information in the synthesized speech.

The speech signal can be processed by the LP analysis to extract the LP residual [7]. The Hilbert envelope (HE) of the LP residual is defined as the magnitude of complex time function (CTF) of LP residual and was first proposed in [8]. The peaks in the HE of LP residual can be detected as epochs [8] [9]. The Group Delay (GD) analysis of LP residual is proposed for the detection of epochs [4]. The later methods include Dynamic Programming Phase Slope Algorithm (DYPSA) [5] and zero frequency filtering (ZFF) [6]. The ZFF method provides best performance in terms of detecting epochs and also computational complexity [6]. The speech signal can be processed by ZFF to find out epochs. The instants of significant excitation can be obtained by anchoring epochs from ZFF method and selecting small range of values around them and also few other instants of excitation that have significant amplitude like those around GOIs.

The present work explores different methods for obtaining the instants of significant excitation. The basic method is to use the sequence of strength weighted epochs detected from the ZFF itself as the instants of significant excitation. It is to be noted that, all the detected epochs, both in the unvoiced and voiced regions are uniformly processed without any distinction. A small range (about 1 ms on either side) of residual samples around the epochs are then used as instants of significant excitation. This is followed by small range of HE of LP residual samples around the epochs as the instants of significant excitation. The fourth method includes using cosine of phase of CTF of LP residual along with the HE of LP residual as the instants of significant excitation. The decomposition of LP residual into HE and cosine of phase gives additional flexibility from the source modeling perspective.

The instants of significant excitation from each of these methods are used as excitation source signal for speech synthesis. The synthesized speech signals are evaluated subjectively for the preservation of naturalness, intelligibility and prosodic information. Based on this, conclusions are made about the significance of instants of significant excitation for source modeling. The rest of the paper is organized as follows: The method for extraction of epochs are described in Section 2. The proposed methods of deriving instants of significant excitation are described in Section 3. The experimental studies, results and discussion are given in Section 4. The summary, conclusion and scope of present work are given in Section 5.

2. Epoch Extraction from Speech

This section briefly reviews the epoch extraction from speech signal using the GD, DYPSA and ZFF methods. In these methods, the epochs refer to glottal closure and glottal opening instants in case of voiced speech, and random excitations like onset of burst and frication in case of unvoiced speech [4].

2.1. Determining Epoch using the GD

The GD method estimates the epochs using the group delay function [4]. First LP residual of speech signal is calculated and then DFT of LP residual and its time weighted version are considered to find phase slope function. The positive zero crossings in the phase slope function give the epochs location.

2.2. Determining Epoch using the DYPSA

The DYPSA method estimates epochs by the identification of peaks in the LP residual of speech using phase slope function [5]. First phase slope function of LP residual is obtained using group delay analysis and the positive zero crossings of phase slope function are hypothesized as the epochs location. The missing epochs are identified using phase slope projection technique which leads to some additional spurious epochs. The spurious epochs are eliminated using the dynamic programming technique.

2.3. Determining Epoch using the ZFF

Epochs can be obtained from the ZFF as follows [6]

- Difference the input speech signal x(n)=s(n)-s(n-1).
- Pass the difference signal x(n) to resonator at 0 Hz twice.

$$y_1(n) = \sum_{k=1}^{4} a_k y_1(n-k) + x(n) \qquad (1)$$

  where a_1=4, a_2=-6, a_3=4 and a_4=-1. This is equivalent to four time successive integration.

- Remove the trend by subtracting y_1(n) with the average value of y_1(n) calculated over window length of average pitch period. The resulting signal is trend removed i.e.,

$$y(n) = y_1(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y_1(n+m) \qquad (2)$$

  where 2N + 1 corresponds to average pitch period.

- The trend removed signal y(n) is termed as ZFF Signal (ZFFS).
- The positive zero crossings of the ZFF signal will give epochs.

Figure 1(a) shows a segment of voiced speech. The processed version of the same using any of the above mentioned epoch extraction methods is shown in Figure 1(b) and corresponds to ZFFS in case of ZFF and phase slope function in case of GD and DYPSA methods. The positive zero crossings of this function are identified as epochs and plotted in Figure 1(c). In case of voiced speech, the epochs occur periodically due to the periodic nature of glottal vibration.

Figure 2(a) shows a segment of unvoiced speech. ZFFS (or phase slope) of the speech signal is shown in Figure 2(b). The positive zero crossings of this function are identified as epochs and plotted in Figure 1(c). In case of unvoiced speech, the epochs occur in random fashion due to the non-periodic nature of excitation.

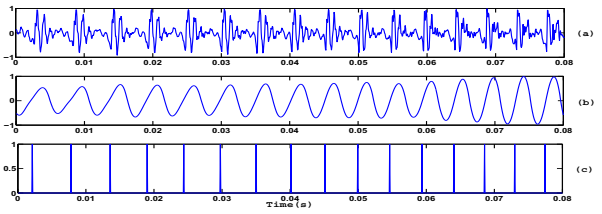

Figure 1: Segment of voiced speech: (a) speech signal, (b) ZFF signal and (c) epoch locations.

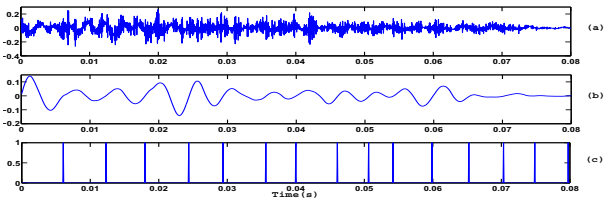

Figure 2: Segment of unvoiced speech: (a) speech signal, (b) ZFF signal and (c) epoch locations.

Even though all the three methods described above are popular for epoch extraction, the ZFF method is shown to give the best performance compared to other methods [6]. This work therefore uses ZFF method used for epoch extraction.

2.4. Strength of Excitation of Epochs

The strength of Excitation ($s_e(ep)$) of epochs is defined as the slope of the ZFF signal [10],

$$s_e(ep) = abs(y(ep+1) - y(ep-1)), \qquad (3)$$

where ep is the epoch location. $s_e(ep)$ gives the strength of impulse-like excitation at the epoch location.

3. Extraction of Instants of Significant Excitation from Speech

In this section different methods to derive the instants of significance excitation are described.

3.1. Using ZFF

In this method, the epochs with their strength calculated by ZFF method used as instants of significant excitation. The epochs are located for the given speech signal and their strength of excitation are computed. The epochs with the corresponding strength of excitation form the instants of significant excitation.

Figure 3(a) shows a segment of speech containing portions of voiced and unvoiced region. Figure 3(b) shows the corresponding residual and Figure 3(c) shows the epochs extracted by the ZFF method. Figure 3(d) shows the epochs with its strength form the instants of significant excitation. The excitation signal is non-zero only at the epochs location.

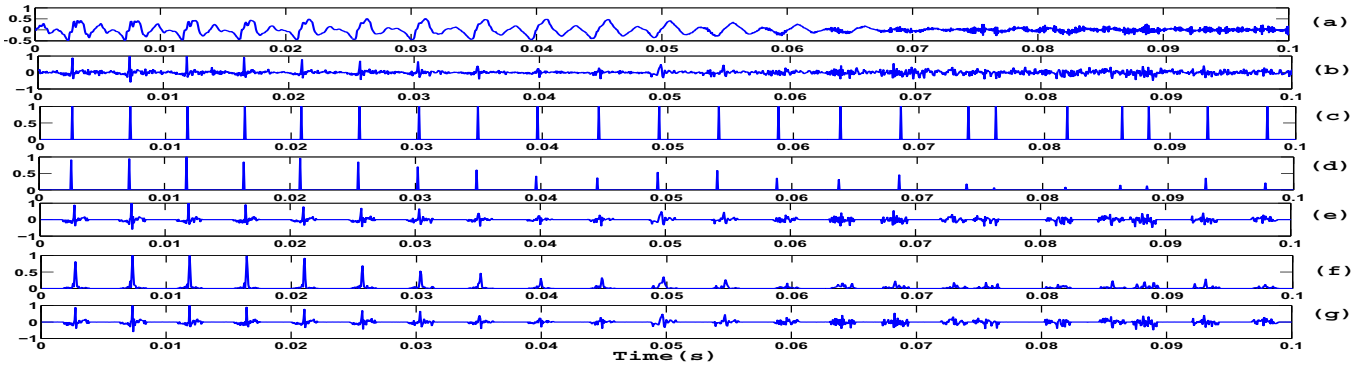

Figure 3: Instants of significant excitations derived from different method: (a) voiced and unvoiced speech portions, (b) LP residual, (c) epochs location calculated from ZFFS, (d) epochs with their strength, instants of significant excitation derived from (e) LP residual, (f) HE and (g) HE+cos_p respectively.

3.2. Using LP residual

In this method, the epochs detected by ZFF method are used as anchor points and a small range of residual samples are extracted around them to form instants of significant excitation. The present work extracts 1 ms of residual samples on either side of the detected epochs.

Figure 3(e) shows the residual extracted by considering 1 ms range on either side of the epochs. The number of non-zero values are more and represent the instants of significant excitation. As a result the source signal in this case may provide significantly better naturalness and preserve the prosodic information.

3.3. Using HE of LP residual

To make the residual samples suitable for modeling, the residual can be further divided into two parts using the definition of CTF. The magnitude of CTF of residual is termed as HE and the real part of the time domain phase of CTF is termed as cosine of the phase. Both these components can be independently modeled for deriving the instants of significant excitation.

Let of r_a(n) be the analytic signal or CTF of residual signal r(n).Then,

$$r_a(n) = r(n) + jr_h(n) \tag{4}$$

where r_h(n) is the Hilbert transform of $r(n)$ which is called as CTF.

$$r_h(n) = IDFT(R_h(\omega)) \tag{5}$$

where

$$R_h(\omega) = \begin{cases} +jR(\omega), \omega < 0 \\ -jR(\omega), \omega > 0 \end{cases} \tag{6}$$

and $R(\omega)$ is the DFT of $r(n)$. DFT and IDFT refers to discrete Fourier transform and inverse discrete Fourier transform, respectively.

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

$$h(n) = |r_a(n)| \tag{7}$$

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \tag{8}$$

In this method, the epochs detected by ZFF method are used as anchor points and a small range of HE samples are extracted around them to form instants of significant excitation. The present work extracts 1 ms of HE samples on either side of detected epochs. Figure 3(f) shows the HE extracted by considering 1 ms range on either side of epochs. The number of non-zero values are more and represent the instants of significant excitation as in the case of residual, but are unipolar in nature.

3.4. Using Cosine Phase of CTF

The cosine phase (cos_p) of CTF is defined as

$$\cos_p = r(n)/h(n) \tag{9}$$

where $r(n)$ is LP residual of speech signal and $h(n)$ is HE of LP residual. Figure 3(g) shows the cosine of phase multiplied with HE by considering 1 ms range on either side of epochs. This case preserves mainly the sequence information as opposed to the strength information present in the HE of LP residual. This sequence information may also be important from the naturalness and prosodic information point of view.

4. Experimental Studies and Discussion

To demonstrate the significance of instants of significant excitation, the speech signal sampled at 16 kHz is processed using 20th order LP analysis, 20 ms frame size and 5 ms frame shift. The LPCs representing the vocal tract information and LP residual representing the excitation source information are obtained. The same speech signal is also processed by ZFF to extract the epochs.

In the first study, epoch sequence with their strength is taken as excitation signal and given to the LP filter for obtaining the synthesized speech by overlap and add method [11]. This method of synthesis is similar to two state LP model [1]. But in this case, voiced/unvoiced separation is not happening and synthesized speech is intelligible which conveys message information, but metallic in quality.

In the second study, in order to incorporate naturalness and prosodic information, LP residual is considered. The excitation source signal now contains the sequence of residual samples anchored around epochs as discussed earlier. The informal listening of synthesized speech gave a feel that it is both natural and intelligible, near to that of original speech. This indicates that the instants of significant excitation should include a small range of samples around epochs for increasing naturalness and preserving prosodic information.

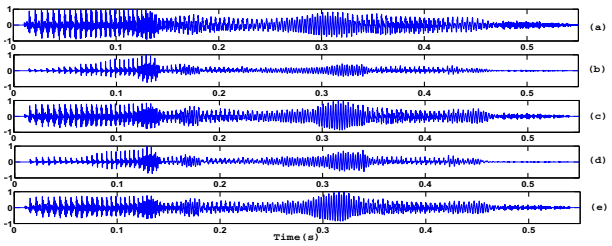

Figure 4: Time domain waveforms of synthesized speech for word *I was*: (a) original speech signal, synthesized speech based on (b) ZFF, (c) LP, (d) HE and (e) HE+cos_p respectively.

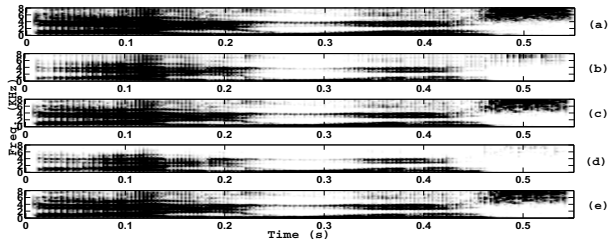

Figure 5: Spectrogram view of synthesized speech for word *I was*: (a) original speech signal, synthesized speech based on (b) ZFF, (d) LP, (c) HE and (e) HE+cos_p respectively.

The motivation for the next two studies is to further understand whether strength or sequence associated with the LP residual is more important for preserving naturalness and intelligibility. The residual is decomposed into HE and cosine of phase representing the strength and sequence information, respectively. The HE values around the epochs are preserved as in the case of residual and used as excitation source signal. The synthesized speech is natural and preserves prosodic information. However, the quality seem to be slightly inferior compared to the residual case. This infers that apart from the strength, sequence information is also important. The synthesis of speech only using cosine of phase resulted in lot of perceptual noise due to the large amplitudes associated with phase sequence values. For this, the instants of significant excitation using HE are multiplied with respective cosine phase values and used as excitation source signal. The synthesized speech quality improves significantly compared using only HE. This study indicates that we need to preserve both strength of excitation and also sequence information in the instants of significant excitation.

Figure 4 and Figure 5 shows the speech waveforms and spectrograms of synthesized speech for the word *I was* by using different instants of significant excitation. It can be seen that there are no discontinuities in the synthesized speech in the case of LP residual and HE+cos_p based system. Most of the features such as pitch changes and formant transitions seems to be preserved well. Hence synthesized speech quality is comparable with the original speech. But in the synthesized speech using HE and ZFF method some discontinuities are seen both in the spectrogram and time domain waveforms, and synthesized speech is not natural, which infers that source of excitation consist of some additional excitation information along with the GCIs, onset of bursts and frication.

4.1. Subjective Study

One sentence *I was about to do this when cooler judgment prevailed* from Arctic database is selected and recorded from 5 speakers (2 male and 3 female) for the study. The recording initially sampled at 48 kHz is down sampled to 16 kHz and used for synthesizing in four different instants of significant excitation cases as explained above. 15 research scholars of our lab participated in the subjective evaluation. The synthesized speech files using all four methods along with the original speech files are presented to the subjects for the evaluations. The speech files were randomized and file names were coded before presenting to the subjects for evaluation. A pilot test was given to each subject before the evaluation. The subjects were asked to observe the naturalness, intelligibility and perceptual distortions present in each file and give there opinion scores accordingly on a standard mean opinion score (MOS) test [12]. There are 25 (5*4 + 5 original files) files used for the evaluation. The mean of the scores obtained for all the files for a given instants of significant excitation technique is calculated as the MOS. The MOS obtained for all the 4 techniques are given in Table 1. We can observe from the Table 1 that there is a significant improvement in MOS scores for the LP residual and HE+cos_p based source models as compared to other two methods. The synthesized files can be accessed from the following link: http://www.iitg.ernet.in/cseweb/tts/Assamese/sourcemodeling.php

Table 1: Mean opinion scores for different source modeling using instants of significant excitation

Modeling technique	MOS
ZFF	2.42
LP	4.05
HE	2.87
HE+cos_p	4.01

5. Summary and Conclusions

In this work, different methods for deriving the instants of significant excitation for source modeling is proposed. Instants of significant excitation are calculated using the epochs from ZFF method as anchor points. The experimental studies indicate that a small set of samples around the epochs are sufficient to preserve the naturalness and prosodic information. Further, both the strength of excitation and sequence knowledge are important from the point of preserving the naturalness and prosodic information. The feasibility of amount of instants of significant excitation required for source modeling has been explored. The future focus may be on source modeling methodologies using instants of significant excitation.

6. Acknowledgements

This work is part of the ongoing project on the development of Text-to-Speech Synthesis for Assamese and Manipuri languages funded by the Technology Development for Indian Languages (TDIL) Program initiated by the Department of Electronics and Information Technology (DeitY), Ministry of Communication and Information Technology, Govt. of India under the consortium mode headed by IIT Madras.

7. References

[1] A.Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, 1994.

[2] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f0 extraction," *Speech Communication*, vol. 27(3-4), pp. 187–207, 1999.

[3] R. Maia, T. Toda, H. Zen, Y. Nankaku, and Tokuda.T, "An excitation model for hmm-based speech synthesis based on residual modeling," in *In Proc. 6th ISCA Workshop Speech Synth.*, 2007.

[4] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech, Audio Proc.*, vol. 3(5), pp. 325–333, 1995.

[5] P. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15(1), pp. 34–43, 2007.

[6] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Audio,Speech, And Language Processing*, vol. 16, pp. 1602–1613, November 2008.

[7] J. Makhoul, "Linear prediction:a tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[8] 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. 27, pp. 309–319, 1979.

[9] K. Sreenivasa Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *Signal Processing Letters, IEEE*, vol. 14(10), pp. 762–765, 2007.

[10] K. S. R. Murthy, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal processing letters*, vol. 16, no. 6, pp. 469–472, June 2009.

[11] R. Crochiere, "A weighted overlap-add method of short time fourier analysis/synthesis," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 1, p. 99102, February 1980.

[12] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication*, vol. 16(3), pp. 225 – 244, 1995.